Neural Network Prediction of Censorable Language

Kei Yin Ng Anna Feldman Jing Peng Chris Leberknight
Montclair State University
Montclair, New Jersey, USA
{ngk2,feldmana,pengj,leberknightc}@montclair.edu

Abstract

Internet censorship imposes restrictions on what information can be publicized or viewed on the Internet. According to Freedom House's annual Freedom on the Net report, more than half the world's Internet users now live in a place where the Internet is censored or restricted. China has built the world's most extensive and sophisticated online censorship system. In this paper, we describe a new corpus of censored and uncensored social media tweets from a Chinese microblogging website, Sina Weibo, collected by tracking posts that mention 'sensitive' topics or authored by 'sensitive' users. We use this corpus to build a neural network classifier to predict censorship. Our model performs with a 88.50% accuracy using only linguistic features. We discuss these features in detail and hypothesize that they could potentially be used for censorship circumvention.

1 Introduction

Free flow of information is absolutely necessary for any democratic society. Unfortunately, political censorship exists in many countries, whose governments attempt to conceal or manipulate information to make sure their citizens are unable to read or express views that are contrary to those in power. One such example is Sina Weibo, a Chinese microblogging website. It was launched in 2009 and became the most popular social media platform in China. Sina Weibo has over 431 million monthly active users[1]. In cooperation with the ruling regime, Weibo sets strict control over the content published under its service. According to Zhu et al. (2013), Weibo uses a variety of strategies to target censorable posts, ranging from keyword list filtering to individual user monitoring. Among all posts that are eventually censored, nearly 30% of them are censored within 5–30 minutes, and nearly 90% within 24 hours (Zhu et al., 2013).Research shows that some of the censorship decisions are not necessarily driven by the criticism of the state (King et al., 2013), the presence of controversial topics (Ng et al., 2018a,b), or posts that describe negative events. Rather, censorship is triggered by other factors, such as for example, the collective action potential (King et al., 2013). The goal of this paper is to compare censored and uncensored posts that contain the same sensitive keywords and topics. Using the linguistic features extracted, a neural network model is built to explore whether censorship decision can be deduced from the linguistic characteristics of the posts.

2 Previous Work

There have been significant efforts to develop strategies to detect and evade censorship. Most work, however, focuses on exploiting technological limitations with existing routing protocols (Leberknight et al., 2012; Katti et al., 2005; Levin et al., 2015; McPherson et al., 2016; Weinberg et al., 2012). Research that pays more attention to linguistic properties of online censorship in the context of censorship evasion include, for example, Safaka et al. (2016) who apply linguistic steganography to circumvent censorship. Lee (2016) uses parodic satire to bypass censorship in China and claims that this stylistic device delays and often evades censorship. Hiruncharoenvate et al. (2015) show that the use of homophones of censored keywords on Sina Weibo could help extend the time a Weibo post could remain available online. All these methods rely on a significant amount of human effort to interpret and annotate texts to evaluate the likeliness of censorship, which might not be practical to carry out for common Internet users in real life. There has also been research that uses linguistic and content

[1] https://www.investors.com/news/technology/weibo-reports-first-quarter-earnings/

Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 40–46
Minneapolis, Minnesota, June 6, 2019.

clues to detect censorship. Knockel et al. (2015) and Zhu et al. (2013) propose detection mechanisms to categorize censored content and automatically learn keywords that get censored. King et al. (2013) in turn study the relationship between political criticism and chance of censorship. They come to the conclusion that posts that have a Collective Action Potential get deleted by the censors even if they support the state. Bamman et al. (2012) uncover a set of politically sensitive keywords and find that the presence of some of them in a Weibo blogpost contribute to higher chance of the post being censored. Ng et al. (2018b) also target a set of topics that have been suggested to be sensitive, but unlike Bamman et al. (2012), they cover areas not limited to politics. Ng et al. (2018b) investigate how the textual content as a whole might be relevant to censorship decisions when both the censored and uncensored blogposts include the same sensitive keyword(s).

3 Tracking Censorship

Tracking censorship topics on Weibo is a challenging task due to the transient nature of censored posts and the scarcity of censored data from well-known sources such as FreeWeibo[2] and WeiboScope[3]. The most straightforward way to collect data from a social media platform is to make use of its API. However, Weibo imposes various restrictions on the use of its API[4] such as restricted access to certain endpoints and restricted number of posts returned per request. Above all, Weibo API does not provide any endpoint that allows easy and efficient collection of the target data (posts that contain sensitive keywords) of this paper. Therefore, an alternative method is needed to track censorship for our purpose.

4 Data Collection

4.1 Web Scraping

4.2 Decoding Censorship

According to Zhu et al. (2013), the unique ID of a Weibo post is the key to distinguish whether a post has been censored by Weibo or has been instead removed by the author himself. If a post has been censored by Weibo, querying its unique ID through the API returns an error message of "permission denied" (system-deleted), whereas a user-removed post returns an error message of "the post does not exist" (user-deleted). However, since the Topic Timeline (the data source of our web scraper) can be accessed only on the front-end (i.e. there is no API endpoint associated with it), we rely on both the front-end and the API to identify system- and user-deleted posts. It is not possible to distinguish the two types of deletion by directly querying the unique ID of all scraped posts because, through empirical experimentation, uncensored posts and censored (system-deleted) posts both return the same error message – "permission denied"). Therefore, we need to first check if a post still exists on the front-end, and then send an API request using the unique ID of post that no longer exists to determine whether it has been deleted by the system or the user. The steps to identify censorship status of each post are illustrated in Figure 1. First, we check whether a scraped post is still available through visiting the user interface of each post. This is carried out automatically in a headless browser 2 days after a post is published. If a post has been removed (either by system or by user), the headless browser is redirected to an interface that says "the page doesn't exist"; otherwise, the browser brings us to the original interface that displays the post content. Next, after 14 days, we use the same methods in step 1 to check the posts' status again. This step allows our dataset to include posts that have been removed at a later stage. Finally, we send a follow-up API query using the unique ID of posts that no longer exist on the browser in step 1 and step 2 to determine censorship status using the same decoding techniques proposed by Zhu et al. as described above (2013). Altogether, around 41 thousand posts are collected, in which 952 posts (2.28%) are censored by Weibo.

topic	censored	uncensored
cultural revolution	55	66
human rights	53	71
family planning	14	28
censorship & propaganda	32	56
democracy	119	107
patriotism	70	105
China	186	194
Trump	320	244
Meng Wanzhou	55	76
kindergarten abuse	48	5
Total	**952**	**952**

Table 1: Data collected by scraper for classification

[2]https://freeweibo.com/

[3]http://weiboscope.jmsc.hku.hk/

[4]https://open.weibo.com/wiki/API文档/en

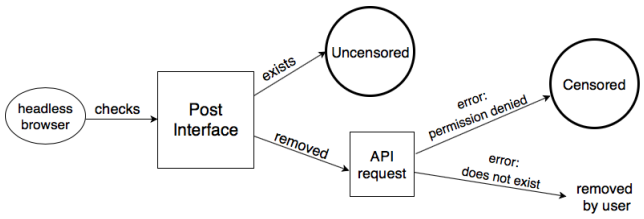

Figure 1: Logical flow to determine censorship status

4.3 Pre-existing Corpus

Zhu et al. (2013) collected over 2 million posts published by a set of around 3,500 sensitive users during a 2-month period in 2012. We extract around 20 thousand text-only posts using 64 keywords across 26 topics (which partially overlap with those of scraped data, see Table 3) and filter all duplicates. Among the extracted posts, 930 (4.63%) are censored by Weibo as verified by Zhu et al. (2013) The extracted data from Zhu et al.(2013)'s are also used in building classification models.

dataset	N	H	features	accuracy
baseline				49.98
human baseline (Ng et al., 2018b)				63.51
scraped	500	50,50,50	Seed 1	80.36
scraped	800	60,60,60	Seed 1	80.2
Zhu et al's	800	50,7	Seed 1	87.63
Zhu et al's	800	30,30	Seed 1	86.18
both	800	60,60,60	Seed 1	75.4
both	500	50,50,50	Seed 1	73.94
scraped	800	30,30,30	all except LIWC	72.95
Zhu et al's	800	60,60,60	all except LIWC	70.64
both	500	40,40,40	all except LIWC	84.67
both	800	20,20,20	all except LIWC	88.50
both	800	30,30,30	all except LIWC	87.04
both	800	50,50,50	all except LIWC	87.24

Table 2: MultilayerPerceptron classification results. N = number of epochs, H = number of nodes in each hidden layer

5 Feature Extraction

We extract features from both our scraped data and Zhu et al.'s dataset. While the datasets we use are different from that of Ng et al. (2018b), some of the features we extract are similar to theirs. We include CRIE features (see below) and the number of followers feature that are not extracted in Ng et al. (2018b)'s work.

topic	censored	uncensored
cultural revolution	19	29
human rights	16	10
family planning	4	4
censorship & propaganda	47	38
democracy	94	53
patriotism	46	30
China	300	458
Bo Xilai	8	8
brainwashing	57	3
emigration	10	11
June 4th	2	5
food & env. safety	14	17
wealth inequality	2	4
protest & revolution	4	5
stability maintenance	66	28
political reform	12	9
territorial dispute	73	75
Dalai Lama	2	2
HK/TW/XJ issues	2	4
political dissidents	2	2
Obama	8	19
USA	62	59
communist party	37	10
freedom	12	11
economic issues	31	37
Total	**930**	**930**

Table 3: Data extracted from Zhu et al. (2013)'s dataset for classification

5.1 Linguistic Features

We extract 4 sets of linguistic features from both datasets – the LIWC features, the CRIE features, the semantics features, and the number of followers feature. We are interested in the LIWC and CRIE features because they are purely linguistic, which aligns with the objective of our study. Also, some of the LIWC features extracted from Ng et al. (2018a)'s data have shown to be useful in classifying censored and uncensored tweets.

LIWC features The English Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2017, 2015) is a program that analyzes text on a word-by-word basis, calculating percentage of words that match each language dimension, e.g., pro-

nouns, function words, social processes, cognitive processes, drives, informal language use etc. LIWC builds on previous research establishing strong links between linguistic patterns and personality/psychological state. We use a version of LIWC developed for Chinese by Huang et al. (2012) to extract the frequency of word categories. Altogether we extract 95 features from LIWC. One important feature of the LIWC lexicon is that categories form a tree structure hierarchy. Some features subsume others.

CRIE features We use the Chinese Readability Index Explorer (CRIE) (Sung et al., 2016), a text analysis tool developed for the simplified and traditional Chinese texts. CRIE outputs 50 linguistic features (see Appendix A.1), such as word, syntax, semantics, and cohesion in each text or produce an aggregated result for a batch of texts. CRIE can also train and categorize texts based on their readability levels. We use the textual-features analysis for our data and derive readability scores for each post in our data. These scores are mainly based on descriptive statistics.

Sentiment features We use BaiduAI[5] to obtain a set of sentiment scores for each post. It outputs a positive sentiment score and a negative sentiment score which sum to 1.

Semantic features We use the Chinese Thesaurus (同义词词林) developed by Mei (1984) and extended by HIT-SCIR[6] to extract semantic features. The structure of this semantic dictionary is similar to WordNet, where words are divided into 12 semantic classes and each word can belong to one or more classes. It can be roughly compared to the concept of word senses. We derive a semantic ambiguity feature by by dividing the number of words in each post by the number of semantic classes in it.

5.1.1 Frequency & readability

We compute the average frequency of characters and words in each post using Da (2004)[7]'s work and Aihanyu's CNCorpus [8] respectively. For words with a frequency lower than 50 in the reference corpus, we count it as 0.0001%. It is intuitive to think that a text with less semantic variety and more common words and characters is relatively easier to read and understand. We derive a Readability feature by taking the mean of character frequency, word frequency and word count to semantic classes described above. It is assumed that the lower the mean of the 3 components, the less readable a text is. In fact, these 3 components are part of Sung et al. (2015)'s readability metric for native speakers on the word level and semantic level.

Followers The number of followers of the author of each post is recorded and used as a feature for classification.

6 Classification

A balanced corpus is created. The uncensored posts of each dataset are randomly sampled to match with the number of their censored counterparts (see Table 1 and Table 3). All numeric values have been standardized before classification. We use the MultilayerPerceptron function of Weka for classification. A number of classification experiments using different combinations of features are carried out. Best performances are achieved using the combination of CRIE, sentiment, semantic, frequency, readability and follower features (i.e. all features but LIWC) (see Table 2).

We also apply the Weka RandomSubset filter using Seed 1 to 8 to randomly select features for classification. The 77 randomly selected features of Seed 1, which is a mix of all features, perform consistently well across the datasets (see Appendix A.1 for the full list of features).

We vary the number of epochs and hidden layers. The rest of the parameters are set to default – learning rate of 0.3, momentum of 0.2, batch size of 100, validation threshold of 20. Classification experiments are performed on 1) both datasets 2) scraped data only 3) Zhu et al.'s data only. Each experiment is validated with 10-fold cross validation. We report the accuracy of each model in Table 2. It is worth mentioning that using the LIWC features only, or the CRIE features only, or all features excluding the CRIE features, or all features except the LIWC and CRIE features all result in poor performance of below 54%.

7 Discussion and Conclusion

Our best results are about 30% higher than the baseline. We also compare our classifiers to the human baseline reported in Ng et al. (2018b). The accuracies of our models are about 25 % higher than the human baseline, which shows that our

[5]https://ai.baidu.com

[6]Harbin Institute of Technology Research Center for Social Computing and Information Retrieval.

[7]http://lingua.mtsu.edu/chinese-computing/statistics/

[8]http://www.aihanyu.org/cncorpus/index.aspx

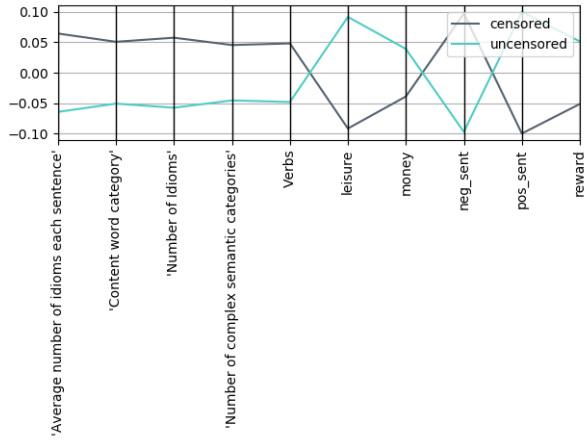

Figure 2: Parallel Coordinate Plots of the top 10 features that have the greatest difference in average values

classifier has a greater censorship predictive ability compared to human judgments. The classification on both datasets together tends to give higher accuracy using at least 3 hidden layers. However, the performance does not improve when adding additional layers (other parameters being the same). Since the two datasets were collected differently and contain different topics, combining them together results in a richer dataset that requires more hidden layers to train a better model. It is worth noting that classifying both datasets using seed 1 features decreases the accuracy, while using all features but LIWC improves the classification performance. The reason for this behavior could be an existence of consistent differences in the LIWC features between the datasets. Since the seed 1 LIWC features (see Appendix A.1) consist of mostly word categories of different genres of vocabulary (i.e. grammar and style agnostic), it might suggest that the two datasets use vocabularies differently. Yet, the high performance obtained excluding the LIWC features shows that the key to distinguishing between censored and uncensored posts seems to be the features related to writing style, readability, sentiment, and semantic complexity of a text.

To gain further insight into what might be the best features that contribute to distinguishing censored and uncensored posts, we compare the mean of each feature of the two classes. The 6 features distinguish censored from uncensored are 1) negative sentiment 2) average number of idioms in each sentence 3) number of idioms 4) number of complex semantic categories 5) verbs 6) number of content word categories. On the other hand, the 4 features that distinguish uncensored from censored are 1) positive sentiment 2) words related to leisure 3) words related to reward 4) words related to money (see Figure 2) This might suggest that the censored posts generally convey more negative sentiment and are more idiomatic and semantically complex in terms of word usage. On the other hand, the uncensored posts might be in general more positive in nature (positive sentiment) and include more content that talks about neutral matters (money, leisure, reward).

To conclude, our work shows that there are linguistic fingerprints of censorship and it is possible to use linguistic properties of a social media post to automatically predict if it is going to be censored. It will be interesting to explore if the same linguistic features can be used to predict censorship on other social media platforms and in other languages.

Acknowledgments

We thank the anonymous reviewers for their careful reading of this article and their many insightful comments and suggestions. This work is supported by the National Science Foundation under Grant No.: 1704113, Division of Computer and Networked Systems, Secure Trustworthy Cyberspace (SaTC).

References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3).

Jun Da. 2004. A corpus-based study of character and bigram frequencies in chinese e-texts and its implications for chinese language instruction. In *The studies on the theory and methodology of the digitalized Chinese teaching to foreigners: Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese.*, pages 501–511. Beijing: Tsinghua University Press.

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*.

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):185–201.

S. Katti, D. Katabi, and K. Puchala. 2005. Slicing the onion: Anonymous routing without pki. Technical report, MIT CSAIL Technical Report 1000.

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):1–18.

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.*

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).

S. Lee. 2016. Surviving online censorship in china: Three satirical tactics and their impact. *China Quarterly*.

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.*

Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. Defeating image obfuscation with deep learning. arXiv preprint arXiv:1609.00408.

jiā jū Méi. 1984. *The Chinese Thesaurus*.

Kei Yin Ng, Anna Feldman, and Chris Leberknight. 2018a. Detecting censorable content on sina weibo: A pilot study. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM.

Kei Yin Ng, Anna Feldman, Jing Peng, and Chris Leberknight. 2018b. Linguistic Characteristics of Censorable Language on SinaWeibo. In *Proceedings of the 1st Workshop on NLP for Internet Freedom held in conjunction with COLING 2018*.

James W. Pennebaker, Roger Booth, and M.E. Francis. 2017. *Linguistic Inquiry and Word Count (LIWC2007)*.

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.

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.*

Yao-Ting Sung, Tao-Hsing Chang, Wei-Chun Lin, Kuan-Sheng Hsieh, and Kuo-En Chang. 2016. Crie: An automated analyzer for chinese texts. *Behavior Research Methods*, 48(4):1238–1251.

Y.T. Sung, T.H. Chang, W.C. Lin, K.S. Hsieh, and K.E. Chang. 2015. Crie: An automated analyzer for chinese texts. *Behavior Research Method*.

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*.

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].

A Appendices

A.1 Appendix I

Full List of CRIE features

*CRIE Readability 1.0
*SVM readability prediction 1.0
Paragraphs
Average paragraph length
*Characters
*Words
Adverbs
*Verbs
Type-token ratio
Difficult words
*Low-stroke characters
*Intermediate-stroke characters
*High-stroke characters
*Average strokes
*Two-character words
*Three-character words
*Sentences
*Average sentence length
*Simple sentence ratio
modifiers per NP
Np ratio
*Average propositional phrase
*Sentences with complex structure
Parallelism
Average number of idioms each sentence
*Content words
*Negatives
*Sentences with complex semantic categories
*Number of complex semantic categories
*Intentional words
*Noun word density
*Content word frequency in logarithmic
*Average frequency of content word in domain in Logarithmic
Number of Idioms
*Pronouns
*Personal pronouns
*First personal pronouns
Third personal pronouns
*conjunctions
positive conjunctions
*negative conjunctions
*adversative conjunctions
*causal conjunctions
hypothesis conjunction
condition conjunction
*purpose conjunctions
*figure of speech (simile)
*Content word category

*feature that is included in Seed 1

Seed 1 LIWC features

WC
WPS
persconc
ppron
we
shehe
they
ipron
quanunit
specart
focuspast
progm
modal pa
general pa
interrog
quant
anx
family
friend
female
differ
see
feel
sexual
drives
achieve
power
risk
motion
work
home
netspeak
assent
Comma
Colon
Exclam
Parenth

Seed 1 semantic, sentiment, and follower features

neg sent
char freq
wc over semantic classes
readability
followers