Chinese Language under Censorship

Shifei Chen
5LN710, Natural Languages Processing, Uppsala University

1 Introduction

In 2017, 64% of netizens lived in countries where their freedom of expression is supressed and for the third consecutive year, China was the world's worst abuser of internet freedom[4]. The censorship system in China behaves in several ways including limited access to websites and keywords, removal of posts on social network websites and blogs, cyber attacks on websites and createing contents for public opinions manipulation. The system works behind th curtain so its organization and mechanism are still not clear to the public. Censored contents usually contain information about pornography, violence and last but least, information against the communist regime. But this doesn't mean all negative information will be censored eventually, there are plenty of them that are considered to be "safe" being expressed freely on the internet. Also, some of the contents that are obfuscated intentially can escape. To understand what kind of expression are more likely to be supressed from a linguistic prospective and how to avoid it by encoding languages, I have conducted this survey base on the First Workshop on Natural Language Processing for Internet Freedom[1].

2 Characteristics of Sensitive Words Blacklists

Censorship in China isn't implemented by the governments or its subordinaries. Service providers are required to comply with content regulations by themselves, which means tech companies and individual developers are liable for the content generated by their users. A simple solution is filtering unfavored information by using a balcklist, which isn't a novelty to block out contents and it is being widely carried out by tech companies and governments around the world, such as Google[2] and IWF[6].

Reverse engineering revealed some blacklists from popular Chinese mobile phone applicaitons[7][10] and games[9]. These studies suggest that, "rather than Chinese censorship being top-down and monolithic, it is a decentralized system where developers are liable for deciding what to censor themselves"[8] since the blacklists found have very little overlap between each other.

If it can be inferred that these big companies have compiled their own blacklists of sensitive words independently since they have the resource to do so, then individual developers are supposed to have a common blacklist as its length are often at tens of thousands level. To find out that, [8] scrapped suspected blacklist files from Chinese open source projects hosted on Github and used machine-learning techniques to determine whether they are blacklists or not.

2.1 Scrapping Blacklists and Extract Their Contents

They used 21924 unique keywords which were found blacklisted by some popular apps and games, combined them with some other common blacklist related words appeared in the file name or the content like "dirty", "sensitive" and "forbid", all of which were observed from previous works, and searched them together on Github to get possible blacklist keywords files from all of the open source projects.

Then they developed a string extractor to extract lists of strings from the file. It supports most-chosen file formats for blacklists by developers such as XML, JSON and CSV, plus C-like source code and plaintexts. However it didn't support SQL database files as their grammar is not consistent.

2.2 Blacklists Classification

To futher classify if the list is a true Chinese blacklist, they used both anaive approach and a machine-learning approach. In the naive approach, researchers simply counts any list that contains "法轮"(Falun Gong) as a blacklist as this word appears in every blacklist publicly known. The limitation of the naive approach is also obvious since it can only find blacklists containing "法轮". Even

though by far it appears in every blacklist, it can not be a guranteened feature of every possbile Chinese blacklist.

In the machine-learning approach, they "used a one-class support vector machine (SVM)[12], as implemented by Scikit-learn[11] and LIBSVM[3]."[8] The classifier requires to be trained by only one class yet maintain the ability to classify both the positive and the negetive results of that class. The training data was the Chinese blacklists that were reverse engineered from popular apps, plus a blacklist from Google[5].

They modelled their possbile blacklists to be "vectors of counts of the number of occurrences of each Chinese word in that list."[8] In addition, to make the calculation tractable, they decided to apply singular-value decomposition to reduce the dimensinality of the dataset. After comparing with the golden standard, which is their result from the naive approach, the final dimension is 46.

2.3 Result

After manual verification, the naive approach returned 884 Chinese blacklists while the machine-learning approach found 1054. The longest blacklist had 38237 words and the mean length and the median length was 2128 and 1026, respectivelly.

These 1054 blacklists contains a wide range of words including pornography, Falun Gong reference, possible independence movements, criticisms to the government, political leader names, etc.

The inter-similarity between blacklists, measured by Jaccard similarity and $max(\%$ of x in y, $\%$ of y in $x)$ for list x and y, were very low. It indicates that there was very little overlap between each blacklist. It corresponds to the same discovery from the blacklists in popular Chinese applicaitons and games. Although how so many individual developers can compose blacklists on their own independently remains unknown, again, both the result from this study and the previous ones had proved that the censorship system in China is a decentralized one.

3 Characteristics of Censorable Language

References

[1] Chris Brew, Anna Feldman, and Chris Leberknight. "Proceedings of the First Workshop on Natural Language Processing for Internet Freedom". In: Proceedings of the First Workshop on Natural Language Processing for Internet Freedom. 2018.

[2] Censorship by Google - Wikipedia. [Online; accessed 6. Jan. 2019]. 2019. URL: https://en.wikipedia.org/wiki/Censorship_by_Google.

[3] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines". In: ACM transactions on intelligent systems and technology (TIST) 2.3 (2011), p. 27.

[4] Freedom on the Net 2017: Manipulating Social Media to Undermine Democracy. [Online; accessed 6. Jan. 2019]. 2017. URL: https://freedomhouse.org/report/freedom-net/freedom-net-2017.

[5] GOOGLE 收集的 GFW 屏蔽关键词（敏感词）. [Online; accessed 7. Jan. 2019]. 2012. URL: https://caiguanhao.wordpress.com/2012/06/01/google-gfw-blacklist.

[6] IWF Response to Prime Minister's Statement. [Online; accessed 6. Jan. 2019]. 2019. URL: https://www.iwf.org.uk/news/iwf-response-to-prime-ministers-statement.

[7] Jeffrey Knockel, Jedidiah R Crandall, and Jared Saia. "Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance." In: FOCI. 2011.

[8] Jeffrey Knockel, Masashi Crete-Nishihata, and Lotus Ruan. "The effect of information controls on developers in China: An analysis of censorship in Chinese open source projects". In: Proceedings of the First Workshop on Natural Language Processing for Internet Freedom. 2018, pp. 1–11.

[9] Jeffrey Knockel, Lotus Ruan, and Masashi Crete-Nishihata. Measuring Decentralization of Chinese Keyword Censorship via Mobile Games. Munk School of Global Affairs, 2017.

[10] Jeffrey Knockel et al. “Every rose has its thorn: Censorship and surveillance on social video platforms in china”. In:

[11] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: Journal of machine learning research 12.Oct (2011), pp. 2825–2830.

[12] Bernhard Schölkopf et al. “Estimating the support of a high-dimensional distribution”. In: Neural computation 13.7 (2001), pp. 1443–1471.