Predicting Censored Content

Chris Leberknight

leberknightc@montclair.edu

Introduction

The following proposal presents details for extending our paper titled, *Leveraging NLP and Social Network Analytic Techniques to Detect Censored Keywords: System Design and Experiments*, which was published at the 52nd Annual Hawaii International Conference on System Sciences in January 2019 (**Track**: Digital Government, **Minitrack**: Dark Digital Government: Exploring the Dangers —Issues, Concerns, and Negative Impacts) [1].

Contribution to Social Computing: Internet regulation and especially Internet censorship has been dramatically increasing since the 1990's. One possible explanation for the increase in online censorship is the idea that the Internet serves as a "democratic enabler," and in many non-democratic countries around the world this poses a significant threat to the status quo. Consequently, information production and consumption in these societies are heavily regulated. However, we also see various levels of censorship employed in democratic nations and frequent national debates on Internet regulation. Another explanation for online censorship is to control or manipulate foreign trade. Foreign corporations that do not abide by domestic censorship polices are legally restricted from conducting business or voluntarily opt to cease operations [4]. Ultimately and regardless of the rationale for the rise of online censorship, citizens in many countries around the world are often fined or imprisoned for attempts to exercise freedom of conscience, press. or expression. Therefore, it is imperative that we raise awareness and gain a deeper understanding of the evolving nature of online censorship to aid in the future development of circumvention tools. It is our hope that this work will inspire novel software tools and polices that enable citizens around the world to openly and freely express their ideas and opinions.

Rationale: Our prior work studies the variability of Internet censorship and automatically categorizes the type of blocked content across different search engines based in mainland China [1]. While our experiments suggest that political content is most frequently blocked compared to other types of content, and the level of censorship varies across search engines, we find these results have three main limitations. First, the list of keywords used in the study is small and may not accurately reflect the type of content being blocked. Second, prior research does not provide insight into how censorship of keywords or content evolves over time. Also, longitudinal studies

that examine censorship from different vantage points are scant [5]. Third, results do not shed light on methods to predict future content that may get censored. The main focus of the proposed research is to investigate a method for predicting censored content.

The key research question that underpins the motivation for the proposed research is:

Can controversy be used a proxy for predicting censored content?

While controversy has been investigated in prior research the methods employed often rely on application specific features [6], [7]. Also, to the best of our knowledge research has yet to examine the correlation between controversial content and censored content. Therefore, we aim to develop a model and metric for measuring controversy in a document that is independent of any feature specific characteristics.

Investigating our main research question will help dissect the censorship process and inform development of novel circumvention tools. For example, if we can show that there is a strong correlation between controversial content and censored content than it would be possible to automatically transform emotionally charged terms or other linguistic properties so that communication succeeds undetected. Exploring this research question may also help identify potential targets used for computational propaganda or disseminating disinformation

Summary for extending prior work: We propose to extend our prior work by developing **(1)** *a probabilistic model and controversy metric for predicting censored content.* We conjecture that controversial content is more likely to be censored compared to non-controversial content. While we acknowledge that not all censored content is controversial, we feel that developing a metric to measure controversy will serve as a key indicator and proxy for predicting censored content. Therefore, we hypothesize that controversial content is more likely to be censored compared to non-controversial content.

Proposal

Our first step to test this hypothesis is to develop a metric to measure the degree of controversy for a given document. It is our hope that results from this work will help to promote free and open communication on the Internet for citizens in countries that rigorously enforce Internet censorship.

To validate our approach, we will test our proposed model and metric with the following labeled datasets.

Datasets

- MPQA – contains 355 news articles from 187 different foreign and U.S. news sources, ranging June 2001 to May 2002 (http://www.cs.pitt.edu/mpqa) [2]
- Wikipedia – Wikipedia maintains a list of over 23,000 controversial articles [3] (https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues)
- FreeWeibo – We have collected over 5,000 censored Weibo posts. FreeWeibo frequently scans and stores censored posts on Weibo (China's version of Twitter)

Research Method (Probabilistic Model): We propose to build a probabilistic model of controversy based on three linguistic analytical methods. We set the algorithmic processes used in controversy detection on a firm theoretical foundation. This first includes creating a specific probabilistic definition of controversy. Then analyze several methods of controversy detection, distilling their essential elements, and combining their strengths together through Bayesian methods. By joining the essential elements together in this way, the most important parts of each method contribute to the whole appropriately. Then we test the performance of our model on data with known labels. Specifically, there are three core methods of extracting relevant information from text.

1. **Sentimentality analysis**: For this, we use *SentiWordNet* (SWN), which returns a positive, negative and objective value for each word. This allows us to assess the emotional content of a document.
2. **Word vectorization:** For this, we use Word2Vec (W2V), which returns a vector of distances from a target word to many other words. This allows us to proportionally determine the degree of relatedness from a target concept to other concepts.
3. **Linguistic analysis:** For this, we use *Linguistic Inquiry and Word Count* (LWIC), which returns a vector of 93 linguistic qualities from a document. This allows us to discover important linguistic properties.

We propose that we take these one at a time and using statistical and machine learning techniques determine the specific qualities that separate controversial vs not controversial text. We then synthesize the data into a probabilistic model that incorporates the important features. Base rates will be estimated and taken into account. The final product will be a single algorithm that will be able to take a word, phrase or document, and output how controversial it is. Once the model is built, we evaluate performance by giving it data with known labels and compute accuracy, recall, precision, and other classification metrics. Figure 1 depicts a high-level overview of pipeline for model construction and analysis.

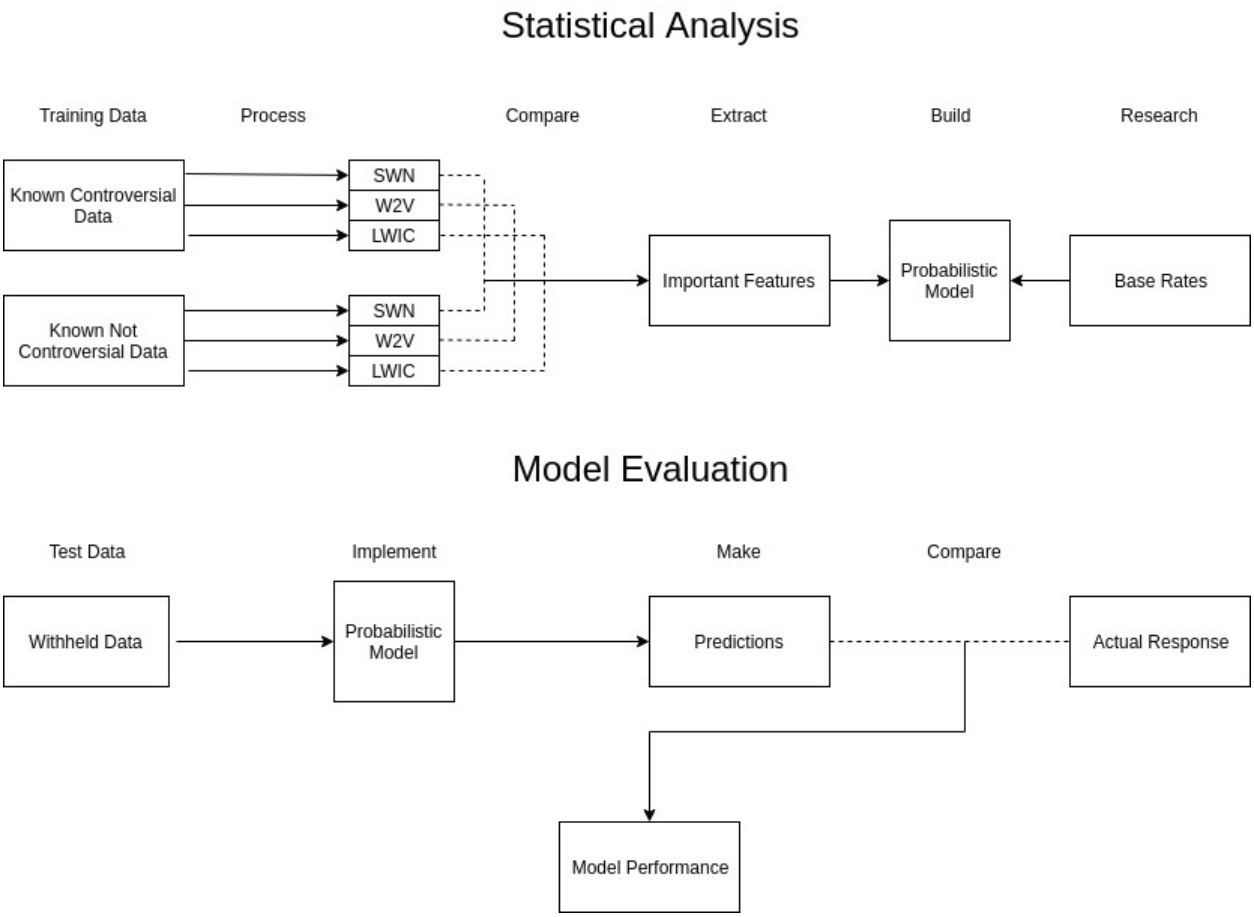

Figure 1: Model Pipeline and Evaluation

We start by considering a word or word phrase *T*. What we ultimately want is $p(Controversial|T)$. The probability that a word or word phrase is controversial. It should be noted that this is different from $p(Controversial, T)$, which would be the probability that a word is controversial and the probability that that phrase is present in the world. It is also different from $p(T|Controversial)$, which is the probability of our word if we already knew it was controversial.

To get $p(Controversial|T)$ we search for *T* in a search engine (preferably in Incognito Mode in order to not bias our results) and collect the first *N* hyper-links. The first *N* hyperlinks form an Order Statistic from our search engine's algorithm of relevance to our search phrase *T*. We form our probabilities from these hyperlinks. It should be noted that when we do this, adding another word phrase to our search term will automatically take into account the probabilistic dependence between two phrases. The search engines algorithm will always display an Order Statistic of the relevance to our phrase no matter how long or short *T* is.

Each hyperlink we call a document D_d. Within each D_d is a set of words W_w. We do analysis on these words in order to determine $p(Controversial|T)$. Currently, we have three core algorithmic methods in order to determine controversy. They are:

- *Algorithm A*: SentiWordNet on each single word W_w
- *Algorithm B*: Word2Vec on each single word W_w
- *Algorithm C*: LWIC on each document D_d

The method we propose blends each of them together as a mixture model with three hyper parameters that need to be determined empirically. It is entirely possible that one or more of these algorithms should be removed and not used. For parsimony, if any of the hyper parameters are close enough to zero, we will remove its corresponding algorithm from our method.

Model: How a model is written down is vitally important to the transmission of its structure. We begin with the blending of the three models $M = \{A, B, C\}$, Where the three p(m) are the three blending hyper parameters discussed above.

$$p(Controversial|T) = \sum_{m \in M} p(Controversial|Tm)p(m) \quad (1)$$

References

[1] Leberknight, C., & Feldman, A. (2019, January). Leveraging NLP and Social Network Analytic Techniques to Detect Censored Keywords: System Design and Experiments. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[2] Choi, Y., Jung, Y., & Myaeng, S. H. (2010, June). Identifying controversial issues and their sub-topics in news articles. In Pacific-Asia Workshop on Intelligence and Security Informatics (pp. 140-153). Springer, Berlin, Heidelberg.

[3] Jankowski-Lorek, Michal & Zielinski, Kazimierz. (2015). Document controversy classification based on the wikipedia category structure. Computer Science. 16. 185. 10.7494/csci.2015.16.2.185.

[4] Leberknight, C. S., Chiang, M., & Wong, F. M. F. (2012). A taxonomy of censors and anti-censors: Part I-impacts of internet censorship. *International Journal of E-Politics (IJEP)*, *3*(2), 52-64.

[5] Burnett, S., & Feamster, N. (2015). Encore: Lightweight measurement of web censorship with cross-origin requests. *ACM SIGCOMM Computer Communication Review*, *45*(4), 653-667.

[6] Shiri Dori-Hacohen and James Allan. 2015. Automated Controversy Detection on the Web. *Advances in Information Retrieval 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015.* 423 (2015).

[7] Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, *1*(1), 3.