Unsupervised Text Classification With Chinese Social Media

An Extension of King, Pan, and Roberts (2017)

Gabe Walker

May 12, 2019

Gov 1006: Models

Abstract

This paper replicates and extends "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument" by G. King, Pan, and Roberts (2017b). It uses an updated version of the readme software ("readme2") by Jerzak, King, and Strezhnev (2019) in conjunction with latent Dirichlet allocation (LDA), a method for unsupervised text classification, to test readme2's performance on an automatically classified set of Chinese social media posts. Using the replication data provided on the Harvard Dataverse[1] and the original five-topic classification scheme, it finds strong support for the authors' original conclusions. It also proposes a new approach for automatically estimating classifications in large corpora when total classification is infeasible with certain machine learning methods.

I. Introduction

The "50 Cent Party" (五毛党) is the commonly used collective term for "internet commentators" (网络评论员), individuals who post pro-government content across the Chinese web.[2] Their activities form an important part of the Chinese government's "public opinion guidance" system, a holistic approach to information control that spans all forms of print, television, and digital media. In the era of WeChat and Weibo, with around 1.5 billion members between the two, public opinion guidance is more important than ever. Over the past decade, journalists, scholars, and the public have closely followed the activities of the 50 Cent Party as a way to understand the Chinese government's approach to managing the country's domestic internet. But because of the opaque and diffuse nature of the Party's operations, it has been a perennial challenge to describe the specific methods that these commentators use.

In "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument," G. King, Pan, and Roberts (2017b) conducted the first large-scale empirical analysis of the 50 Cent Party's operations. Using nearly 44,000 leaked

[1] G. King, Pan, and Roberts (2017a)

[2] The name "50 Cent" comes from the widely held belief that commentators get paid 0.50 RMB per post.

web posts from the Zhanggong Internet Propaganda Office in Jiangxi Province, they characterize the timing and content of 50 Cent Party activity on the site. They then extrapolate their findings to China as a whole. Overall, they conclude that "prevailing views" about the 50 Cent Party are "largely incorrect." Previously, the majority of journalists and academics believed that Party members posted nationalistic and pro-government content and argumentatively tackled controversial debates head-on. King, Pan, and Roberts find, however, that these commentators often post positive, "cheerleading" content that avoids touchy subjects and outright criticism. They also find that 50 Cent commentators coordinate the timing and content of their activities and probably publish around 450 million posts a year across the Chinese web.

This paper aims to replicate King, Pan, and Roberts's results and to employ their methods in a new way. Specifically, I use readme2, an improved software for proportion estimation by Jerzak, King, and Strezhnev (2019) to reproduce their results, and a fully automated clustering approach in concert with readme2 to "automate" proportion estimation. I find that my results align closely with those reported in the original paper and that there may be interesting potential to use readme2 to estimate proportions in large document corpora without human assistance.

II. Replication

In order to replicate some of the results in G. King, Pan, and Roberts (2017b), I obtained the paper's original datasets of collected posts from the Harvard Dataverse. The first set of interest is the 43,757 known 50 Cent Party posts mentioned in the leaked files from the Zhanggong Internet Propaganda Office. These posts date from 2013 and 2014 and appear on a wide variety of different Chinese websites, including social media, discussion forums, and government-run sites. More than half of those on commercial sites appeared on Weibo. The second set of interest is scraped Weibo posts from "exclusive" sources named in the Zhanggong leak: that is, a collection of 5,584 posts from accounts that almost never post anything *besides* 50 Cent Party content. I chose these two accounts to attempt to replicate a portion of the original findings. The one difference in my approach is that I use readme2, and updated version of the readme software the authors used in their 2017 publication.

First, I show that the timing of the known posts matches up exactly with what the authors reported. See the appendix for the reproduced original figure (Figure 2) from the paper.

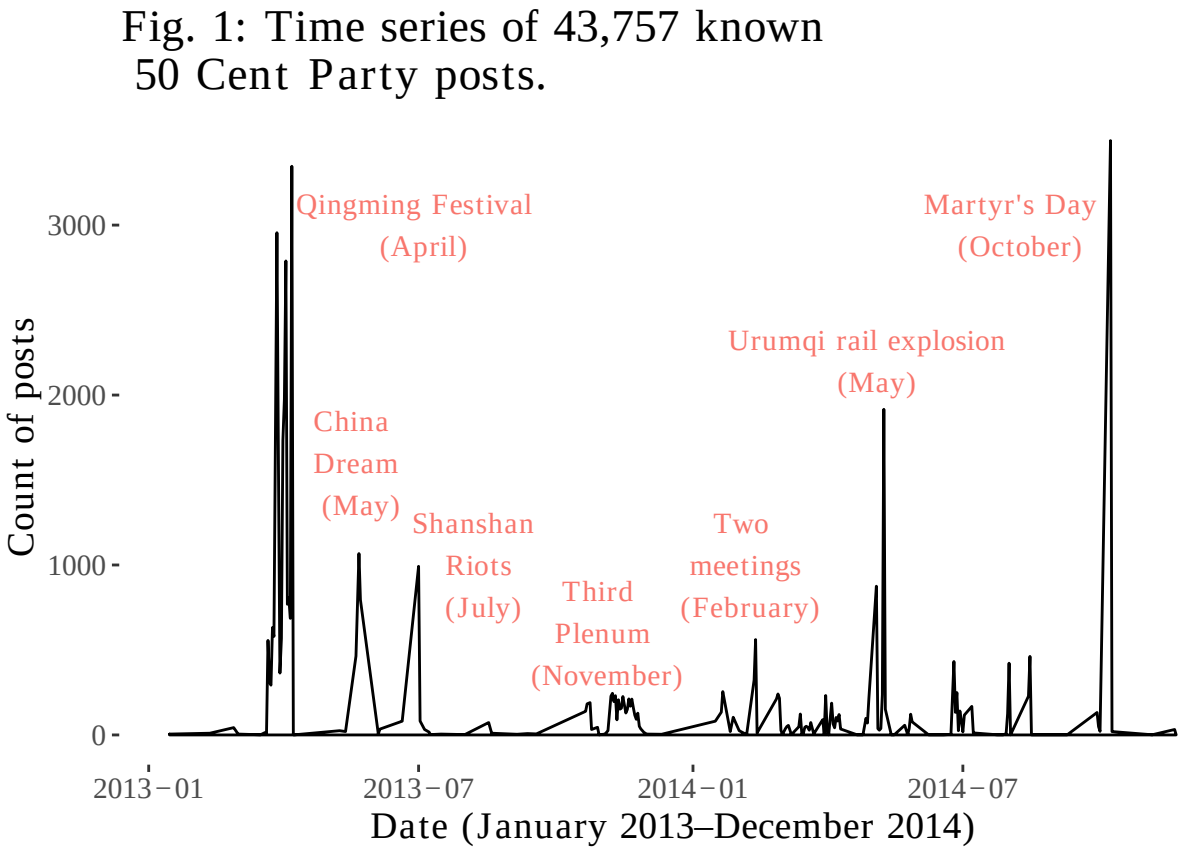


Figure 1: The peaks show counts of identified 50 Cent Party posts during significant events (noted in red) during 2013–2014.

Next, I use readme2 to estimate the proportions of the different categories in the first set of 43,757 posts. Both readme and readme2 work as described in Hopkins et al. (2012) by using human-determined categories to estimate the proportions of those categories in the entire corpus. For example, in this case the authors initially chose five possible categories of posts.[3] Then two Chinese speakers were provided with 200 randomly selected posts (of the 43,757) to sort. The 188 that they agreed upon are then fed into readme, which provides an unbiased estimate of the proportions of each category in the *entire* collection of posts. The key distinction of this approach compared to many other methods of text classification is that readme does *not* actually classify anything. readme2 operates in the same way with some new improvements as described in Jerzak, King, and Strezhnev (2019). The process for estimating proportions for the exclusive posts is exactly the same.

[3]Taunting of foreign countries; argumentative praise or criticism; nonargumentative praise or criticism; factual reporting; cheerleading for China.

The mean results of three runs of readme2 on the original datasets (leaked and exclusive posts), juxtaposed with results from the original paper, show close alignment between the two.[4]

Fig. 2: The readme2 results match well with those from the original paper.

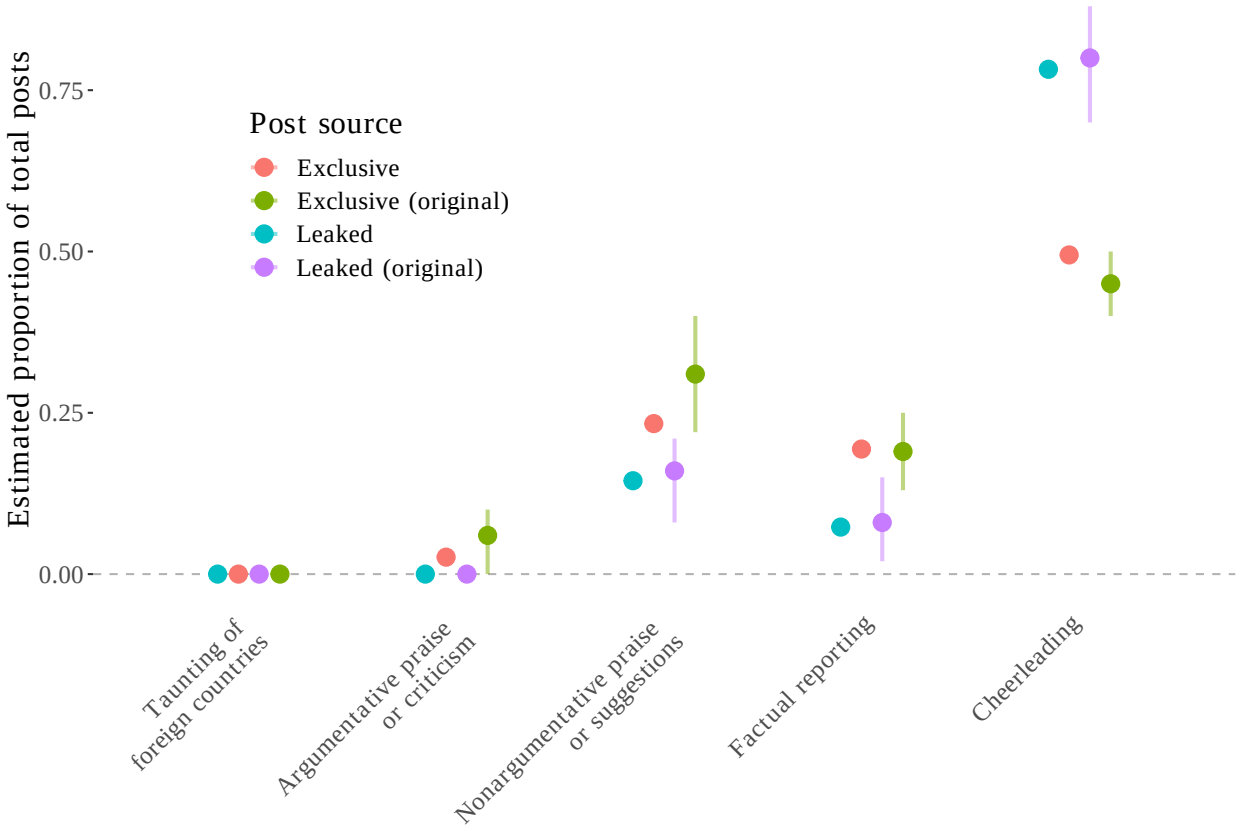


Figure 2: Each color corresponds to a run of readme2. Estimates of the proportions of each of the five categories are given on the vertical axis, with error bars around estimates from the original paper. Leaked posts number 43,757 and posts from "exclusive" 50 Cent Party accounts number 5,584.

As evident in Fig. 2, nearly all of my results match well with those from the original paper. None of the human-coded posts included examples of "taunting of foreign countries," so readme ignores that category completely. The leaked posts had no "argumentative praise or criticism" either—so in this case readme essentially provided estimates of only *three* categories and assumed the others were zero. For the other three categories, my readme2 results and the original results align within the margin of error given in Figure 3 of the paper. A greater number of repeated trials of readme2 would have provided a confidence interval around the mean estimates (i.e., the points on the graph), but the obvious similarity would have remained.

[4]These are results I estimated from Figure 3 of the paper, reproduced in the appendix.

Overall, both results suggest that 50 Cent Party posts in these samples are primarily "cheerleading for China" (between 50 and 75 percent) and "nonargumentative praise or suggestions" and "factual reporting" (around 30–50 percent together). Very few, if any, are "argumentative" or "taunting of foreign countries."

III. Extension

A few aspects of G. King, Pan, and Roberts (2017b) prompted me to apply their methods in a new way. First, the paper's choice of five categories seemed somewhat arbitrary. For example, this post is an example of one classified as "argumentative praise or criticism":

> Lee Kai-Fu says that you can buy a villa for $600,000 USD in New York, much cheaper than in Beijing. But what he doesn't tell you is that this so-called villa is actually a warehouse, which is more than a four-hour drive from New York City.
>
> 李开复说纽约 60 万美金一套别野，比北京便宜多了，但他不胡告诉你那套所谓的别野其实是个仓库，而且离纽约市区车需要四个多小时。

This post clearly aligns with the authors' definition because it includes criticism of an individual. But it also seems to taunt the United States and may even have elements of factual reporting. What would happen if we defined starting categories that were completely different than these five? And would readme faithfully return unbiased estimates of these new categories?

In the first step of my extension I implement a different approach to classifying the known 50 Cent Party posts. Rather than initially hand-coding a subset of posts I use latent Dirichlet allocation (LDA), just one of more than 150 fully automated classification techniques reported in Grimmer and King (2011), to cluster posts. To do this, I selected a subset of 5,000 of the 43,757 (to save computation time) to transform into a "document term matrix," a numeric array with the frequencies of every word contained in the corpus. After eliminating sparse terms (that appeared in fewer than 1 in every 1,000 posts) and sparse posts (that did not contain any common terms), I applied a basic LDA model without any specified hyperparameters to identify five distinct topics.[5]

[5]Only to keep some general equivalence to the original paper. Note that because the objective functions of different automated clustering approaches are totally different, this LDA outcome is just one out of many possible classification schemes that we might expect to see. And five categories is in this case an arbitrary choice without any measures of purity within each group.

The table below contains the top eight terms of each of five categories, my names for the categories, and the proportion of each in the 5,000 posts.

Topic	Name	Proportion	Translation	Terms
1	Local Progress	0.20	Ganzhou, the masses, hope, Southern Ganzhou, government, development, two, establish	赣州, 群众, 希望, 赣南, 政府, 发展, 两, 建设
2	National Development	0.15	China, reform, dream, development, society, realize, economy, Party	中国, 改革, 梦, 发展, 社会, 实现, 经济, 党
3	Ancestral Memory	0.14	Qingming, Qingming Festival, martyr, heroic martyr, cherish, give thanks, memorialize, civilized	清明, 清明节, 先烈, 英烈, 缅怀, 感恩, 祭奠, 文明
4	Red Spirit (present)	0.26	revolution, the people, martyr, lucky, the homeland, life, hero, martyr	革命, 人民, 烈士, 幸福, 祖国, 生活, 英雄, 先烈
5	Red Spirit (past)	0.25	martyr, revolution, spirit, soviet, cherish, revitalize, development, Southern Ganzhou	先烈, 革命, 精神, 苏区, 缅怀, 振兴, 发展, 赣南

As is clear from the top terms, distinguishing between each category is not easy. This could be in part due to the nature of LDA, which assigns posts to multiple topics and words to multiple topics; that is, there is natural cross-listing between topics and key words. In this case, a few categories share "martyr," "revolution," and "development." It is also possible that five categories is a poor number for this subset of the known 50 Cent Party posts; a greater number might differentiate between topics more clearly. I also attempted the same analysis with a biterm LDA—using pairs of adjacent words—but the result was no more distinct. Regardless of the final real-world significance of these categories, the algorithm did successfully sort the posts, which lays the groundwork for the next step of my extension.

To test readme2's performance I ran 45 trials of readme2 "seeded" with 188 randomly selected categorized posts from the 5,000 subset. Again, this number is somewhat arbitrary, and could be increased for greater accuracy. The results show that readme2 estimates accurate proportions of the LDA classifications in the overall document.

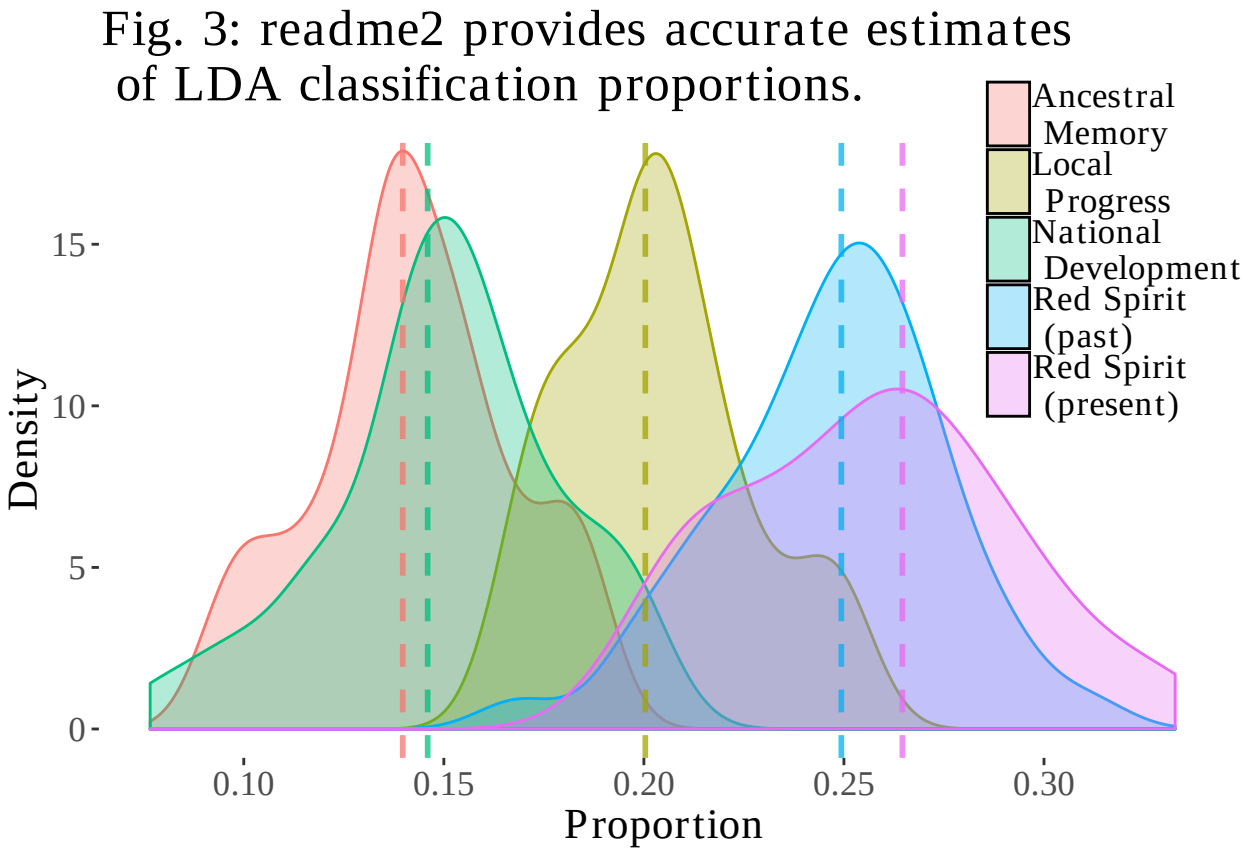


Figure 3: The colored distributions are estimates of the proportions of each group in the 5,000-post subset from 45 readme2 runs. The dotted lines are the actual proportions as determined by LDA.

Fig. 3 shows that readme2 provides excellent estimates of the actual proportions that LDA classified (shown with the dotted lines). It even captures the difference between "national development" and "ancestral memory," which differ by only around 1 percent. However, though the means are not statistically different (with a one-sample t-test), there is some variance to the individual readme outputs, likely based on the initial "seed" proportions. For example, readme2's maximum and minimum estimates for "Red Spirit (present)" are 31 and 17 percent, respectively, with a mean of 24.7 percent (actual is 24.9).

In one respect this test of readme2 is a meaningless exercise: LDA already classified all of the subsample posts so we do not need to know the "actual" proportions. In another respect, however, it may have future usefulness. Some unsupervised classification methods, such as hierarchical clustering on a high-dimensional document feature matrix, are challenging to compute on some machines. The above results suggest that readme2 may provide a

way to "extrapolate" proportions from a small set to a large set. For example, the unsupervised approach could label a small sample set, "seed" readme2 with those, and then have readme2 repeatedly estimate the proportions in the much larger dataset. This could provide a useful workaround for directly classifying each and every document in a large corpus and an interesting exploratory step for automatically surveying the landscape of an expansive collection of documents.

Acknowledgements

Many thanks to David Kane and Mark Hill for their enthusiasm and support, and to Gary King for his helpful suggestions.

Appendix: Replicated Figures

Figure 2: Time Series of 43,757 Known 50c Social Media Posts with Qualitative Summaries of the Content of Volume Bursts

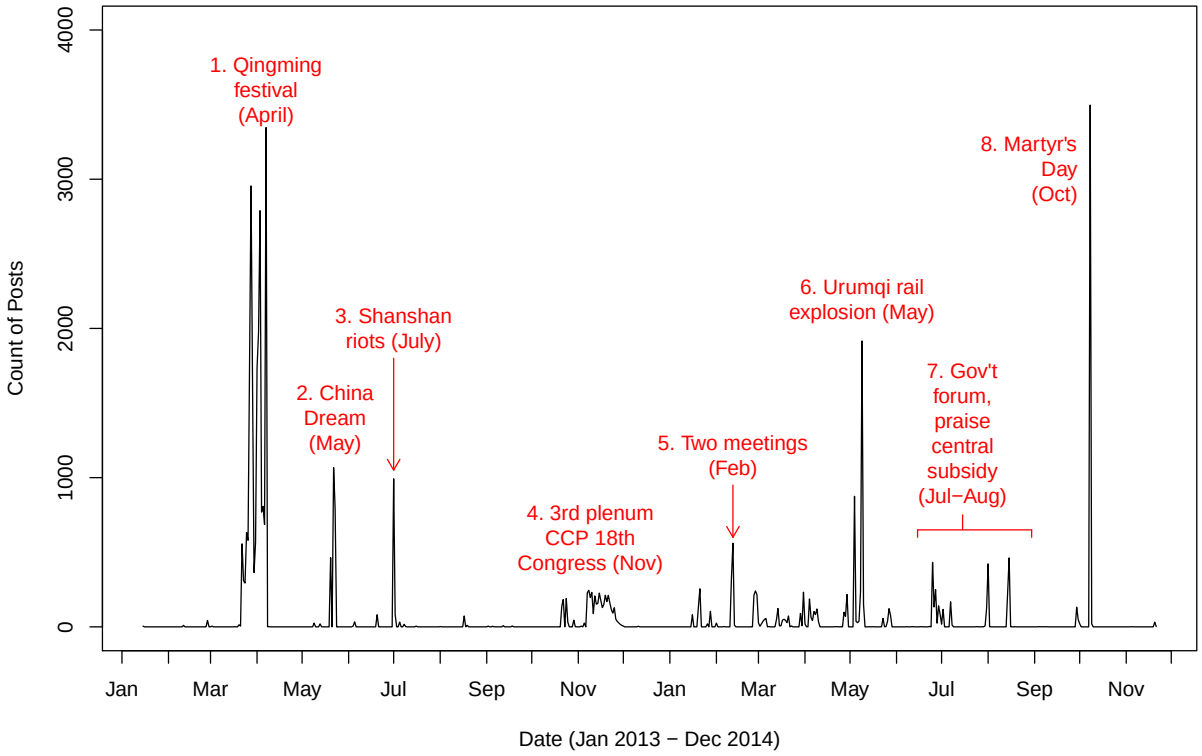
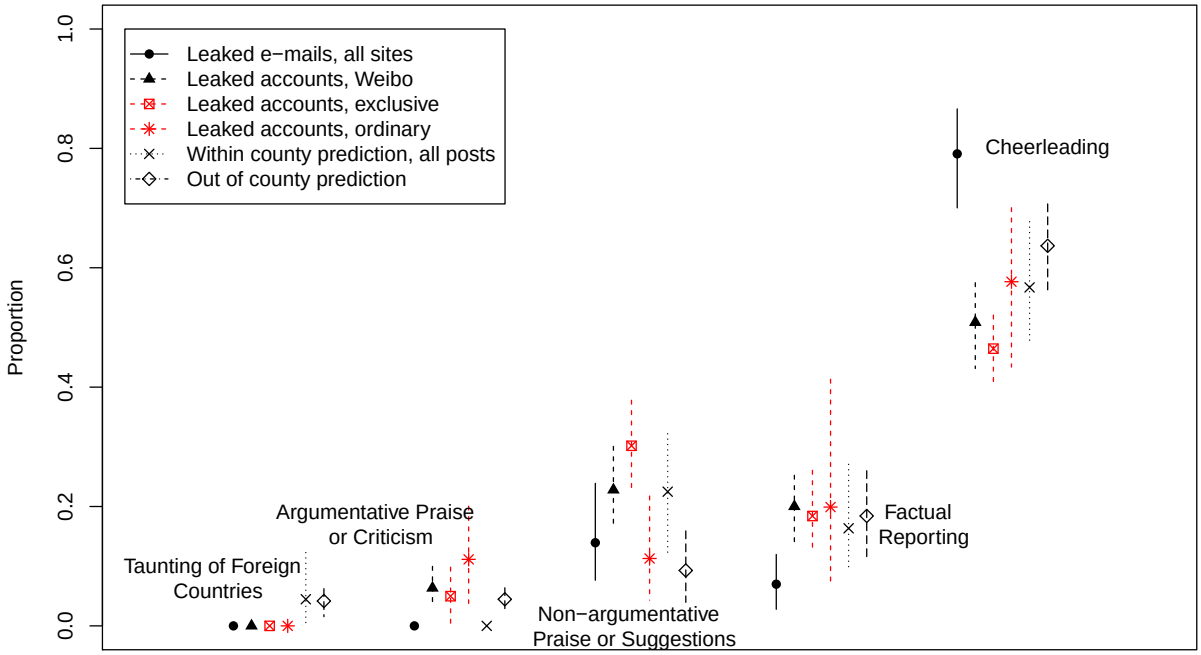


Figure 3. Content of Leaked and Inferred 50c Posts, by Substantive Category



References

Grimmer, J., and G. King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108 (7): 2643–50. https://doi.org/10.1073/pnas.1018067108.

Hopkins, Daniel, Gary King, Matthew Knowles, and Steven Melendez. 2012. "ReadMe: Software for Automated Content Analysis." http://gking.harvard.edu/files/gking/files/readme.pdf.

Jerzak, Connor T., Gary King, and Anton Strezhnev. 2019. "An Improved Method of Automated Nonparametric Content Analysis for Social Science," Working Paper,. https://gking.harvard.edu/files/gking/files/word.pdf.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2017a. "Replication data for: How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." Harvard Dataverse. https://doi.org/10.7910/DVN/QSZMPD.

———. 2017b. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111 (3): 484–501. https://doi.org/10.1017/S0003055417000144.