Censorship of Online Encyclopedias: Implications for NLP Models

Eddie Yang*
z5yang@ucsd.edu
University of California, San Diego
La Jolla, California

Margaret E. Roberts*
meroberts@ucsd.edu
University of California, San Diego
La Jolla, California

ABSTRACT

While artificial intelligence provides the backbone for many tools people use around the world, recent work has brought to attention that the algorithms powering AI are not free of politics, stereotypes, and bias. While most work in this area has focused on the ways in which AI can exacerbate existing inequalities and discrimination, very little work has studied how governments actively shape training data. We describe how censorship has affected the development of Wikipedia corpuses, text data which are regularly used for pre-trained inputs into NLP algorithms. We show that word embeddings trained on Baidu Baike, an online Chinese encyclopedia, have very different associations between adjectives and a range of concepts about democracy, freedom, collective action, equality, and people and historical events in China than its regularly blocked but uncensored counterpart – Chinese language Wikipedia. We examine the implications of these discrepancies by studying their use in downstream AI applications. Our paper shows how government repression, censorship, and self-censorship may impact training data and the applications that draw from them.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Information systems** → **Content analysis and feature selection**; • **Social and professional topics** → **Political speech**.

KEYWORDS

word embeddings, censorship, training data, machine learning

ACM Reference Format:
Eddie Yang and Margaret E. Roberts. 2021. Censorship of Online Encyclopedias: Implications for NLP Models. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445916

1 INTRODUCTION

Natural language processing (NLP) as a branch of artificial intelligence provides the basis for many tools people around the world use daily. NLP impacts how firms provide products to users, content individuals receive through search and social media, and how individuals interact with news and emails. Despite the growing importance of NLP algorithms in shaping our lives, recently scholars, policymakers, and the business community have raised the alarm of how gender and racial biases may be baked into these algorithms. Because they are trained on human data, the algorithms themselves can replicate implicit and explicit human biases and aggravate discrimination [6, 8, 39]. Additionally, training data that over-represents a subset of the population may do a worse job at predicting outcomes for other groups in the population [13]. When these algorithms are used in real world applications, they can perpetuate inequalities and cause real harm.

While most of the work in this area has focused on bias and discrimination, we bring to light another way in which NLP may be affected by the institutions that impact the data that they feed off of. We describe how censorship has affected the development of online encyclopedia corpuses that are often used as training data for NLP algorithms. The Chinese government has regularly blocked Chinese language Wikipedia from operating in China, and mainland Chinese Internet users are more likely to use an alternative Wikipedia-like website, Baidu Baike. The institution of censorship has weakened Chinese language Wikipedia, which is now several times smaller than Baidu Baike, and made Baidu Baike - which is subject to pre-censorship - an attractive source of training data. Using methods from the literature on gender discrimination in word embeddings, we show that Chinese word embeddings trained with the same method but separately on these two corpuses reflect the political censorship of these corpuses, treating the concepts of democracy, freedom, collective action, equality, people and historical events in China significantly differently.

After establishing that these two corpuses reflect different word associations, we demonstrate the potential real-world impact of training data politics by using the two sets of word embeddings in a transfer learning task to classify the sentiment of news headlines. We find that models trained on the same data but using different pre-trained word embeddings make significantly different predictions of the valence of headlines containing words pertaining to freedom, democracy, elections, collective action, social control, political figures, the CCP, and historical events. These results suggest that censorship could have downstream effects on AI applications, which merit future research and investigation.

Our paper proceeds as follows. We first describe the background of how Wikipedia corpuses came to be used as training data for word embeddings and how censorship impacts these corpuses. Second, we describe our results of how word associations from Wikipedia and Baidu Baike word embeddings differ on concepts that pertain

*Both authors contributed equally to this research.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445916

to democracy, equality, freedom, collective action and historical people and events in China. Last, we show that these embeddings have downstream implications for AI models using a sentiment prediction task.

2 PRE-TRAINED WORD EMBEDDINGS AND WIKIPEDIA CORPUSES

NLP algorithms rely on numerical representations of text as a basis for modeling the relationship between that text and an outcome. Many NLP algorithms use "word embeddings" to represent text, where each word in a corpus is represented as a k-dimensional vector that encodes the relationship between that word and other words through the distance between them in k-dimensional space. Words that frequently co-occur are closer in space. Popular algorithms such as Glove [30] and Word2Vec [24] are used to estimate embeddings for any given corpus of text. The word embeddings are then used as numerical representations of input texts, which are then related through a statistical classifier to an outcome.

In comparison to other numerical representations of text, word embeddings are useful because they communicate the relationships between words. The bag-of-words representation of text, which represents each word as simply being included or not included in the text, does not encode the relationship between words – each word is equidistant from the other. Word embeddings, on the other hand, communicates to the model which words tend to co-occur, thus providing the model with information that words like "purse" and "handbag" as more likely substitutes than "purse" and "airplane".

Word embeddings are also useful because they can be pre-trained on large corpuses of text like Wikipedia or Common Crawl, and these pre-trained embeddings can then be used as an initial layer in applications that may have less training data. Pre-trained word embeddings have been shown to achieve higher accuracy faster [31]. While training on large corpuses is expensive, companies and research groups have made available pre-trained word embeddings – typically on large corpuses like Wikipedia or Common Crawl – that can then be downloaded and used in any application in that language.[1]

The motivation behind using pre-trained word embeddings is that they can reflect how words are commonly used in a particular language. Indeed, Spirling and Rodriguez [36] show that pre-trained word embeddings do surprisingly well on a "Turing test" where human coders often cannot distinguish between close words produced by the embeddings and those produced by other humans. To this end, Wikipedia corpuses are commonly selected to train word embeddings because they are user-generated, open-source, cover a wide range of topics, and are very large.[2]

At the same time as pre-trained embeddings have become popular for computer scientists in achieving better performance for NLP tasks, some scholars have pointed to potential harms these embeddings could create by encoding existing biases into the representation. The primary concern is that embeddings replicate existing human biases and stereotypes in language and using them in downstream applications can perpetuate these biases (see Sun et al. [37] for a review). Caliskan et al. [8] show that word embeddings reflect human biases, in that associations of words in trained word embeddings mirror implicit association tests. Using simple analogies within word embeddings, Bolukbasi et al. [6], Garg et al. [14], and Manzini et al. [23] show that word embeddings can encode racial and gender stereotypes. While these word associations can be of interest to social science researchers, they may cause harm if used in downstream tasks [3, 29].

More generally, research in machine learning has been criticized for not paying enough attention to the origin of training datasets and the social processes that generate them [15]. Imbalances in the content of training data have been shown to create differential error rates across groups in areas ranging from computer vision to speech recognition [40, 41]. Some scholars have argued that training datasets should be representative of the population that the algorithm is applied to [35].

3 CENSORSHIP OF CHINESE LANGUAGE WIKIPEDIA AND IMPLICATIONS FOR CHINESE LANGAUGE NLP

We consider another mechanism through which institutional and societal forces impact the corpuses that are used to train word embeddings: government censorship. While we use the example of online encyclopedias and word embeddings to make our point, its implications are much more general. Government censorship of social media, news, and websites directly affects large corpuses of text by blocking users' access, deleting individual messages, adding content through propaganda, or inducing self-censorship through intimidation and laws [11, 19, 20, 22, 25, 32, 34].

While Wikipedia's global reach makes it an attractive corpus for training models in many different languages, Wikipedia has also been periodically censored by many governments, including Iran, China, Uzbekistan, and Turkey [10]. China has had the most extensive and long-lasting censorship of Wikipedia. Chinese language Wikipedia has been blocked intermittently ever since it was first established in 2001. Since May 19, 2015, all of Chinese language Wikipedia has been blocked by the Great Firewall of China [27, 44]. More recently, not just Chinese language Wikipedia, but all language versions of Wikipedia have been blocked from mainland China [2].

Censorship has weakened Chinese language Wikipedia by decreasing the size of its audience. Pan and Roberts [28] estimate that the block of Chinese language Wikipedia in 2015 decreased page views of the website by around 3 million views per day. Zhang and Zhu [48] use the 2005 block of Wikipedia to show that the block decreased views of Chinese language Wikipedia, which in turn decreased user contributions to Wikipedia not only from blocked users in mainland China, but also from unblocked users what had fewer incentives to contribute after the block. While mainland Chinese Internet users can access Chinese language Wikipedia with a Virtual Private Network (VPN), evidence suggests that very few do [9, 32].

[1] For example, Facebook's provides word embeddings in 294 languages trained on Wikipedia (https://fasttext.cc/docs/en/pretrained-vectors.html [5].
[2] A Google Scholar search of "pre-trained word embeddings" and Wikipedia returns over 2,000 search results as of January 2021.

Censorship of Chinese language Wikipedia has strengthened its unblocked substitute, Baidu Baike. A similar Wikipedia-like website, Baidu Baike as of 2019 boasted 16 million entries, 16 times larger than Chinese language Wikipedia [46]. Yet, as with all companies operating in China, Baidu Baike is subject to internal censorship that impacts whether and how certain entries are written. While edits to Chinese language Wikipedia pages are posted immediately, any edits to Baidu Baike pages go through pre-publication review. While editors of Wikipedia can be anonymous, editors of Baidu Baike must register their real names. Additional scrutiny is given to sensitive pages, such as national leaders, political figures, political information, and the military, where Baidu Baike regulations stipulate that only government media outlets such as *Xinhua* and *People's Daily* can be used as sources.[3]

Pre-censorship of Baidu Baike affects the types of pages available on Baidu Baike and the way these pages are written. While it's impossible to know without an internal list the extent to which missing pages in Baidu Baike are a direct result of government censorship, a substantial list of historical events covered on Chinese language Wikipedia including "June 4th Incident" and "Democracy Wall" and well-known activists such as Chen Guangcheng and Wu'erkaixi have no Baidu Baike page [26]. For example, when we attempted to create entries on Baidu Baike such as "June Fourth Movement" or "Wu'erkaixi," we were automatically returned an error.

Perhaps because of the size difference between the two corpuses, increasingly researchers developing cutting edge Chinese langauge NLP models are drawing on the Baidu Baike corpus [38, 43]. Baidu Baike word embeddings have been shown to perform better on certain tasks [21]. Here, we assess the downstream implications of this choice on the representation of democratic concepts, social control, and historical events and figures. First, we follow Caliskan et al. [8] to compare the distance between these concepts and a list of adjectives and sentiment words. Then, we show the downstream consequences of the choice of corpus on a predictive task of the sentiment of headlines.

4 DISTANCE FROM DEMOCRACY: COMPARISON BETWEEN BAIDU BAIKE AND WIKIPEDIA EMBEDDINGS

In this section, we consider the differences in word associations among word embeddings trained with Chinese language Wikipedia and Baidu Baike. We use word embeddings made available by Li et al. [21].[4] Li et al. [21] train 300-dimensional word embeddings on both Baidu Baike and Chinese language Wikipedia using the same algorithm, Word2Vec [24]. For a benchmark, we also compare these two sets of embeddings to embeddings trained on articles from the *People's Daily* from 1947-2016, the Chinese government's mouthpiece.[5]

To evaluate word associations, we follow Caliskan et al. [8] and Rodman [33] to compare the distance between a set of target words and attribute words to establish their relationships in each embedding space. Figure 1 gives a simplified graphical representation of the evaluation procedure in a 2-dimensional space. In this simple example, we might be interested in the position of a target word – a concept we are interested in – relative to a positive attribute word and a negative attribute word. For example, we can evaluate whether democratic concepts are represented more positively or negatively by comparing the angle between the vector for the target word "Democracy" (in black) and a positive attribute word "Stability" as well as a negative attribute word "Chaos" (both in blue).

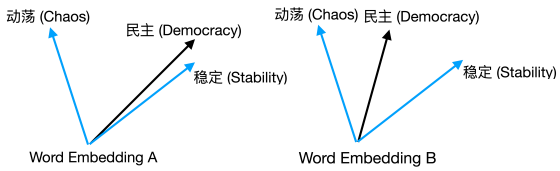

Figure 1: Example of Word Embedding Comparison

In Figure 1, "Democracy" in word embedding A has a more positive connotation than in word embedding B, because the relative position of the word "Democracy" in embedding A with respect to the positive attribute word "Stability" and the negative attribute word "Chaos" is closer to the positive attribute word than "Democracy" is in embedding B. To minimize the particularities of a single word and hence the variability of the result, we repeat this evaluation procedure across multiple target words representing the same concept (e.g. democracy) and compare them with multiple attribute words. In the next sections, we explain how we select target words, attribute words, how we pre-process the embedding space, and our results.

4.1 Identifying Target Words

We begin by delineating the categories of interest. In general, there are two broad categories we are interested in: 1) democratic political concepts and ideas and 2) known targets of propaganda. Based on past work, we know entries that fall under these categories have been the target of content control on Baidu Baike [26].Additionally, the first category captures ideas that we think are normatively desirable but discouraged in China. The second category captures the extent that the embeddings are consistent with propaganda.

For the first category, we include

(1) Democratic values, in particular freedom and equality of rights.
(2) Procedures of democracy, in particular features pertaining to elections.
(3) Channels for voicing preferences in the form of collective actions such as protests and petitions.

For the second category, we include

(1) Social control, especially concepts related to repression and surveillance.
(2) The Chinese Communist Party (CCP) and related features.

[3]See instructions at: https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91%EF%BC%9A%E5%8F%82%E8%80%83%E8%B5%84%E6%96%99.
[4]https://github.com/Embedding/Chinese-Word-Vectors
[5]Also trained by Li et al. [21] and made available at https://github.com/Embedding/Chinese-Word-Vectors.

(3) Significant historical events in China that involved the CCP, such as the Cultural Revolution.
(4) Important figures who are extolled by the CCP.
(5) Figures who are denounced by the CCP, such as political dissenters.

For each of these categories, we do not want to select only one target word of interest, but rather a group of related words that all cover the same concept. We select a group of target words that "represent" this category as follows:

(1) For categories other than historical events and negative figures, we first select a Chinese word that most closely represents the category of interest.[6] For example, for the category of procedures of democracy, the Chinese word "election" is selected.
(2) We then calculate the cosine similarity of the representative word with all other words from the word embedding spaces (Wikipedia & Baidu Baike).
(3) From each corpus, we select 50 words that are closest to the representative word (words with the highest cosine similarity).
(4) Of the 100 words closest to the representative word for each category, we include all words that could be thought to be synonymous or a subset of the more general category. We drop those that are domain specific; for example, of the words for the category of procedures of democracy, we dropped the word "Japanese Diet", which is specific to the Japanese political system.
(5) For categories on historical events and negative figures, we simply used the name of the person or of the historical event.
(6) The full list of words for each category is presented in Appendix D.

We opt for the data-driven approach in (3) and (4) to select target words in order to limit researcher degree of freedom. Furthermore, the selection of representative words in (1) and the pruning of synonyms in (4) were done by three native Chinese speakers to ensure the selected words provide good coverage of how the categories of interest are discussed in the Chinese context.

4.2 Selecting Attribute Words

We use two strategies for selecting attribute words. First, we draw on the literature on propaganda in China to select a set of positive and negative words that would be consistent with what we know about CCP propaganda narratives. As scholars of propaganda have pointed out, the CCP has actively tried to promote the image of itself and China's political system as stable and prosperous, while characterizing Western democratic systems as chaotic and in economic decline [1, 7, 47]. Therefore, for our first set of words, which we call "Propaganda Attributes Words," positive words include synonyms of stability and prosperity, while negative attribute words include synonyms of chaos, decline, and instability. The full list for the set of propaganda attribute words is presented in Appendix E.

For the second set of words, we are interested in whether the target words are more generally evaluated differently between the two corpuses. To test this, we make use of a dictionary of evaluative words specifically designed for Chinese natural language processing [42]. The dictionary codes whether an evaluative word is positive, negative, or neutral. We follow the preprocessing instructions by Wang and Ku [42] by dropping all neutral words and only using the list of positive and negative evaluative words. A sample of the set of evaluative words is presented in Appendix F. For subsequent discussions, we refer to this list of attribute words as the "Evaluative Attribute Words."

[6] We asked three Chinese speakers to independently come up with the representative words and had them agree on a single word for each category. This step was done before analysis was performed.

4.3 Pre-processing Word Embedding Spaces

There are two notable challenges when comparing different word embeddings. One, word embeddings produced by stochastic algorithms such as Word2Vec will embed words in non-aligned spaces defined by different basis vectors. This precludes naive comparison of word distances across distinct corpuses [17, 33]. If the centroids of the two word embeddings are different, then using cosine similarity (i.e. the cosine of the angle between two vectors) to compare word associations across different corpuses can yield uninterpretable result. Figure 2 presents a simplified example of this problem. One word embedding, by virtue of being further away from the origin, yields a smaller angle between the two vectors, even though the relative positions of the two vectors in the two word embeddings are the same.

To solve this problem, we standardize the basis vectors of each word embeddings by subtracting the means and dividing by the standard deviations of the basis vectors, so that each word embedding is centered around the origin with dimension length 1.

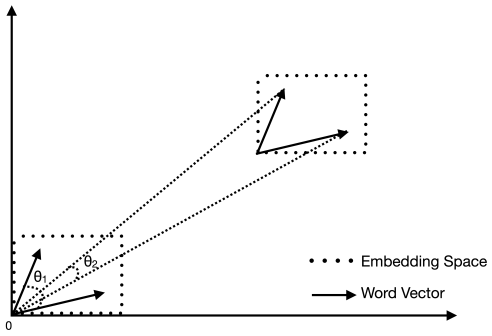

Figure 2: Nonalignment between Two Word Embeddings

Another problem is that word embeddings trained on different corpuses can have different vocabulary. This precludes us from comparing words that appear in one word embedding but are not present in the other word embedding. Because of this, we only keep the intersection of the vocabularies of word embeddings. As a result, six target words were dropped in the comparison between Wikipedia- and Baidu Baike-trained word embeddings and five target words were dropped in the comparison between Wikipedia- and *People's Daily*-trained word embeddings.

4.4 Expectations

We expect ideas that are normatively appealing but discouraged in China to be portrayed more negatively in Baidu Baike. We expect figures who are denounced by the CCP to be portrayed more negatively in Baidu Baike. On the other hand, we expect categories that are targets of positive propaganda to be portrayed more positively in Baidu Baike. Overall, we expect that censorship and curation of Baidu Baike will mean that the words we are interested in will be treated similarly in Baidu Baike and state media outlet *The People's Daily*. A summary of our theoretical expectations is presented in Table 1 below.

Table 1: Theoretical Expections

Category	Sign
Freedom	−
Democracy	−
Election	−
Collective Action	−
Negative Figures	−
Social Control	+
Surveillance	+
CCP	+
Historical Events	+
Positive Figures	+

Note: Negative sign indicates Baidu Baike and *People's Daily* are less favorable than Wikipedia and positive sign indicates that Baidu Baike and *People's Daily* are more favorable than Wikipedia.

4.5 Limitations

Through this design, we test whether there are differences between word embeddings trained on Chinese language Wikipedia and those trained on Baidu Baike in topics where there is evidence of censorship on Baidu Baike. While we think the evidence we produce is suggestive that censorship impacts the placement of the word embeddings, we cannot isolate the effect of censorship outside of other differences that may exist between Baidu Baike and Chinese language Wikipedia. Isolating the effect of censorship is difficult in part because censorship's influence is pervasive, affecting the content not only through pre-publication review, but also likely through the propensity for individuals to become editors and the information that they have and are willing to contribute. This makes it very difficult to establish a counterfactual of what the content on Baidu Baike would have looked like without censorship. We believe Chinese language Wikipedia is the closest approximation to this counterfactual.

4.6 Results

Following Caliskan et al. [8], we use a randomization test with one-sided p-value to compare how words in each category are represented in Wikipedia, Baidu Baike and *People's Daily*.

Formally, let X_i, $i \in a, b$ be the set of word vectors for the target words from embedding a and b respectively. Let A_i, B_i, $i \in a, b$ be the two sets of word vectors for the attribute words, with A being the set of positive attributes and B being the set of negative attributes. Subscript i again denotes the embedding that the word vectors are from. Let $\cos(\vec{p}, \vec{q})$ denote the cosine of the angle between vectors $\vec{p}$ and $\vec{q}$. The test statistic is

$$s_i(X, A, B) = \sum_{i \in a} s(x_i, A_i, B_i) - \sum_{i \in b} s(x_i, A_i, B_i)$$

where

$$s(t, A, B) = \text{mean}_{p \in A} \cos(\vec{t}, \vec{p}) - \text{mean}_{q \in B} \cos(\vec{t}, \vec{q})$$

Let Ω denotes the set of all possible randomization realizations of assignment of word vector x to embedding $i \in \{a, b\}$. The one-sided p-value of the permutation test is

$$\Pr_i[s_{\omega \in \Omega}(X, A, B) > s_i(X, A, B)]$$

We present the effect size of the difference in word associations across word embeddings, defined as

$$\frac{\text{mean}_{i \in a} s(x_i, A_i, B_i) - \text{mean}_{i \in b} s(x_i, A_i, B_i)}{\text{std.dev}_i s(x_i, A_i, B_i)}$$

Conventional cutoffs for small, medium, and large effect sizes are 0.2, 0.5, and 0.8, respectively. The comparisons between Wikipedia and Baidu Baike word embeddings and between Wikipedia and *People's Daily* word embeddings are presented in Table 2 and Table 3 respectively.

Table 2: Wikipedia vs. Baidu Baike

	Propaganda Attributes		Evaluative Attributes	
	effect size	p-value	effect size	p-value
Freedom	-0.62	0.01	0.06	0.60
Democracy	-0.50	0.05	-0.56	0.03
Election	-0.27	0.13	-0.33	0.05
Collective Action	-0.66	0.00	-0.09	0.34
Negative Figures	-0.91	0.00	0.50	0.99
Social Control	0.70	0.04	0.68	0.01
Surveillance	0.09	0.32	0.73	0.00
CCP	1.05	0.02	1.39	0.00
Historical Events	0.14	0.19	0.27	0.01
Positive Figures	0.59	0.00	1.17	0.00

Across most categories and for both sets of attribute words, the differences in word embeddings are in line with our theoretical expectations. Table 2 indicates that for categories Freedom, Democracy, Election, Collective Action, and Negative Figures, word embeddings trained with Baidu Baike display a more negative connotation than embeddings trained with Wikipedia. For categories Social Control, Surveillance, CCP, and Historical Events, word embeddings trained with Baidu Baike display a more positive connotation than embeddings trained with Wikipedia. The effect sizes indicate substantial differences for target words that are related to democracy and those that are targets of propaganda. This is consistent across both set of attribute words and across the two comparisons. In Table 3 we show that the effect sizes when comparing Wikipedia and Baidu Baike are similar to comparing Wikipedia with the government publication *The People's Daily*.

Table 3: Wikipedia vs. *People's Daily*

	Propaganda Attributes		Evaluative Attributes	
	effect size	p-value	effect size	p-value
Freedom	-0.29	0.11	-0.51	0.01
Democracy	-0.40	0.09	-0.97	0.00
Election	-0.43	0.04	-0.91	0.00
Collective Action	-0.81	0.00	-0.10	0.34
Negative Figures	0.44	0.91	-0.06	0.41
Social Control	0.82	0.01	0.58	0.03
Surveillance	0.31	0.06	0.84	0.00
CCP	1.39	0.00	1.22	0.00
Historical Events	0.29	0.08	0.22	0.04
Positive Figures	1.51	0.00	1.29	0.00

While most categories accord with our expectations, one in particular deserves further explanation. Negative figures, including activists and dissidents who the CCP denounces, are only more significantly associated with negative words on Baidu Baike and *People's Daily* in one instance and even have a positive effect size comparing Baidu Baike to Wikipedia in Table 2. It is likely that because of censorship there is very little information about these figures in the Baidu Baike and *People's Daily* corpuses, so their word embeddings do not show strong relationships with the attribute words. To examine this, we used Google Search to count the number of pages on Chinese language Wikipedia and Baidu Baike that link to each negative figure. Out of 18 negative figures, Chinese language Wikipedia has more page links to two thirds of them, even though Chinese language Wikipedia is 16 times smaller. Therefore, the uncertainty around the result we have for negative figures may be a result of lack of information about these individuals in Baidu Baike.

5 APPLICATION: SENTIMENT ANALYSIS OF NEWS HEADLINES

In this section, we demonstrate that the differences we detected in word embeddings have tangible effect on downstream machine learning tasks. To do this, we make use of the pre-trained word embeddings on each of the different corpuses as inputs in a larger machine learning model that automatically labels the sentiment polarity of news headlines. We chose the automated classification of news headlines because machine learning based on news headlines is used in recommendation systems for social media news feeds and news aggregators, as well as for analysts using automated classification of news to make stock price and economic predictions.[7] We show that using the pre-trained word embeddings from Baidu Baike and Chinese language Wikipedia with identical training data produces sentiment predictions for news headlines that differ systematically across our categories of interest.

[7]For example, EquBot https://equbot.com/.

5.1 Data and Method

We imagine a scenario where the task is to label the sentiment of news headlines where the model is trained on a large, general sample of news headlines. We then examine the performance of this model on an oversample of headlines that include our target words. This allows us to evaluate how a general news sentiment classifier performs on words that are politically valanced in China, varying the origin of the pre-trained embeddings, but holding constant the sentiment labels in the training and test sets.

For the training set, we randomly select 5,000 headlines from the TNEWS dataset. The TNEWS dataset contains 73,360 Chinese news headlines of various categories.[8] It is part of the Chinese Language Understanding Evaluation (CLUE) Benchmark and is widely used as the training data for Chinese news classification models. For each of the randomly selected 5,000 headlines, we label each news headline as positive, negative, or neutral in line with the general sentiment of the headline. For our training set from the TNEWS dataset, we have 1,861 headlines with positive sentiment, 781 with negative sentiment, and 2,342 with neutral sentiment.[9]

For the test set, we collect Chinese news headlines that contain any of our target words from Google News. For each of the target words, we collect up to 100 news headlines. Because some target words yield only a handful of news headlines, we collected 12,669 news headlines in total, out of 182 target words. Data collection was done in July and August of 2020. Using the exact same coding scheme as the training set, we label these headlines as positive, negative, or neutral. The test set contains 5,291 headlines with positive sentiment, 3,913 with negative sentiment, and 3,424 with neutral sentiment.[10]

We preprocess the news headlines by removing punctuation, numbers, special characters, the names of the news agency (if they appear on the headline), and duplicated headlines. To convert the news headlines into input for machine learning models, we first use a Chinese word segmentation tool to segment each news headline into a sequence of words. We then look up the word embedding for each word in the sequence. Following a conventional approach, we take the average of the pre-trained word embeddings of the words in a given news headline to represent each headline. Any word that does not have a corresponding word embedding in the Word2Vec models is dropped. This leaves us with three different representations of the headlines: one for Baidu Baike, one for Chinese language Wikipedia, and one for the *People's Daily*.

With each of these three different representations of the text based on different pre-trained embeddings, we train three machine learning models – Naive Bayes (NB), support vector machines (SVM) and TextCNN [18]. For each model, we use identical training labels, from the TNEWS dataset.[11] This yields a total of nine models, with three for each pre-trained word embeddings. Each trained model is then used to predict sentiment labels on the test set. Because of the

[8]For more details about the TNEWS dataset, see Appendix.
[9]16 duplicated news headlines are dropped, resulting in 4,984 headlines in total.
[10]41 duplicated news headlines are dropped, resulting in 12,628 headlines in total.
[11]Because headlines with neutral labels are more noisy and given the difficulty of training a three-class classifier with limited training data, we report results in the main text based on models that are trained with only positive and negative headlines. We report results with neutral headlines included in the Appendix. Our substantive conclusions are largely intact.

stochastic nature of TextCNN, the TextCNN results are averaged over 10 runs for each model.

We compare different trained models of the same architecture (NB, SVM, or TextCNN) by looking at the mis-classifications for each category of target words. Intuitively, a model that is pre-disposed to associate more positive words with a certain category of headlines will have more false-positives (e.g. negative headlines mis-classified as positive), whereas a model that is pre-disposed to associate more negative words with a certain category of headlines will have more false-negatives (e.g. positive headlines mis-classified as negative).

Because the overall mis-classification rate may differ for headlines of different target words, we use a linear mixed effects model to compare the different embeddings, allowing headlines of different target words to have different intercepts. More formally, let L_{ij} be a list of N human-labeled sentiment scores for headlines containing target word i in category j. Let $\hat{L}^a_{ij}$ and $\hat{L}^b_{ij}$ be the predicted sentiment scores from model a and b for the same headlines. We estimate the linear mixed effects model for each category j of news headlines by

$$y_j = \alpha_{ij} + X_j\beta_j + \epsilon_j \quad (1)$$

where the outcome variable y_j is a $2N \times 1$ vector of difference in classifications against human labels, $\begin{matrix}\hat{L}^a_j - L_j \\ \hat{L}^b_j - L_j\end{matrix} \sum \alpha_{ij}$ is a $2N \times 1$ vector of random intercepts corresponding to headlines of each target word i in category j. X_j is an indicator variable for model a (as opposed to b) and β_j is the coefficient of interest.

5.2 Results

Before turning to the results of the impact of pre-trained embeddings on the predicted classifications of the model, we report the overall accuracy of each of the models on the test set in Table 4. Overall, TextCNN performs the best out of the three models. However, within models no set of pre-trained word embeddings performs better than the other – they all perform quite similarly.

Table 4: Model Accuracy in Test Set

	Model	Accuracy
Naive Bayes		
	Baidu Baike	76.83
	Wikipedia	76.29
SVM		
	Baidu Baike	77.12
	Wikipedia	76.68
TextCNN		
	Baidu Baike	82.84
	Wikipedia	81.60

Even though the selection of pre-trained embeddings does not seem to impact overall accuracy, the pre-trained embeddings do influence the false positive and false negative rates of different categories of headlines. In Table 5 we show the comparison of Baidu Baike and Wikipedia, where Baidu Baike is model a and Wikipedia is model b. This means X_j from Equation 1 is 1 for category j if the model were trained with Baidu Baike word embeddings and 0 for Wikipedia. A negative coefficient indicates that on average Baidu Baike rates this category more negatively than Wikipedia. A positive coefficient indicates that on average Baidu Baike rates this category as more positive than Wikipedia.

Table 5: Baidu Baike vs. Wikipedia

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.13	0.00	-0.06	0.00	-0.04	0.04
Democracy	-0.08	0.00	-0.05	0.04	-0.04	0.06
Election	-0.11	0.00	-0.06	0.03	-0.02	0.48
Collective Action	-0.13	0.00	-0.07	0.00	-0.05	0.01
Negative Figures	-0.04	0.03	0.00	0.96	-0.01	0.54
Social Control	0.03	0.12	0.00	0.93	0.03	0.13
Surveillance	-0.01	0.68	-0.01	0.80	0.00	0.91
CCP	0.03	0.21	0.01	0.65	0.03	0.05
Historical Events	-0.04	0.04	0.01	0.75	-0.02	0.26
Positive Figures	0.06	0.00	0.06	0.00	0.06	0.00

The results are largely consistent with what we found in Section 4. Overwhelmingly, Wikipedia predicts headlines that contain target words in the categories of freedom, democracy, election, and collective action to be more positive. In contrast, Baidu Baike predicts headlines that contain target words of figures that the CCP views positively to be more positive. The exceptions to our expectations are the categories of social control, surveillance, CCP, and historical events, where we cannot reject the null of no difference between the two corpuses, although they do not go against our expectations. We find similar results for the comparison between *People's Daily* and Chinese language Wikipedia, in Table 6.

Table 6: *People's Daily* vs. Wikipedia

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.22	0.00	-0.08	0.00	-0.12	0.00
Democracy	-0.14	0.00	-0.06	0.02	-0.07	0.00
Election	-0.13	0.00	-0.01	0.62	-0.04	0.12
Collective Action	-0.19	0.00	-0.05	0.05	-0.06	0.00
Negative Figures	0.01	0.78	0.01	0.72	-0.05	0.01
Social Control	0.05	0.00	0.01	0.66	0.01	0.63
Surveillance	-0.04	0.11	-0.02	0.34	-0.03	0.22
CCP	0.07	0.00	0.00	0.82	0.02	0.24
Historical Events	-0.01	0.77	0.02	0.29	-0.01	0.44
Positive Figures	0.13	0.00	0.04	0.00	0.06	0.00

To provide intuition, Figure 3 shows examples of headlines labeled differently between model trained with Baidu Baike pre-trained embeddings and model trained with Chinese language Wikipedia in our test set. The model trained with Baidu Baike pre-trained word embedding labeled "Tsai Ing-wen: Hope Hong Kong Can Enjoy Democracy as Taiwan Does" as negative, while Wikipedia and humans labeled this headline as positive. The difference in these predictions do not stem from the training data – which is the same

– or the model – which is the same. Instead, the associations made within the pre-trained word embeddings drive these differences.

Example 1: 蔡英文: 盼台湾享有的民主自由香港也可以有
Tsai Ing-wen: Hope Hong Kong Can Enjoy Democracy as Taiwan Does
Baidu Baike Label: - Wikipedia Label: + Human Label: +

Example 2: 封杀文化席卷欧美 自由反被自由误?
Cancel Culture Spreading through the Western World, Is It the Fault of Freedom?
Baidu Baike Label: - Wikipedia Label: + Human Label: -

Example 3: 共产暴政录: 抗美援朝真相
Communist Tyranny: The Truth about Chinese Involvement in the Korean War
Baidu Baike Label: + Wikipedia Label: - Human Label: -

Example 4: 香港《国安法》: 中国驻港部队司令强硬表态维稳
Hong Kong Security Law: PLA Hong Kong Garrison Commander Takes Tough Stance in Support of Stability Maintenance
Baidu Baike Label: + Wikipedia Label: - Human Label: -

Figure 3: Examples of Headlines Labeled Differently By Naive Bayes Models Trained with Baidu Baike and Wikipedia

6 CONCLUSION

The extensive use of censorship in China means that the Chinese government is in the dominant position to shape the political content of large Chinese language corpuses. Even though corpuses like Chinese language Wikipedia exist outside of the Great Firewall, they are significantly weakened by censorship, as shown by the smaller size of Chinese language Wikipedia in comparison to Baidu Baike. While more work would need to be done to understand how these discrepancies affects users of any particular application, we showed in this paper that political differences reflective of censorship exist between two of the corpuses commonly used to train Chinese language NLP. While our work focuses on word embeddings, the discrepancies we uncovered likely affect other pre-trained NLP models as well, such as BERT [12] and ERNIE [38]. Furthermore, these political differences present a pathway through which political censorship can have downstream effects on applications that may not themselves be political but that rely on NLP, from predictive text and article recommendation systems to social media news feeds and algorithms that flag disinformation.

The literature in computer science has taken on the problem of bias in training data by looking for ways to de-bias it – for example, through data augmentation [49], de-biasing word embeddings [6], and adversarial learning [45].[12] However, it is unclear how to think about de-biasing attitudes toward democracy, freedom, surveillance, and social control. What does unbiased look like in these circumstances, and how would one test it? The only way we can think about an unbiased training set in this circumstance is one where certain ideas are not automatically precluded from being included in any given corpus. But knowing what perspectives have been omitted is difficult to determine and correct after the fact.

[12] Although methods for de-biasing have also been shown to often be inadequate [4, 16].

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation under Grant No.˜0001738411. Thanks to Guanwei Hu, Yucong Li, and Zoey Jialu Xu for their excellent research assistance. We thank Michelle Torres, Allan Dafoe, and Jeffrey Ding for their helpful comments on this work.

REFERENCES

[1] 2016. No News is Bad News. *The Economist* (2016). https://www.economist.com/china/2016/02/04/no-news-is-bad-news
[2] 2019. Wikipedia blocked in China in All Languages. *BBC News* (2019). https://www.bbc.com/news/technology-48269608
[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv preprint arXiv:2005.14050* (2020).
[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.
[7] Anne-Marie Brady. 2015. Authoritarianism Goes Global (II): China's Foreign Propaganda Machine. *Journal of Democracy* 26, 4 (2015), 51–59.
[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[9] Yuyu Chen and David Y. Yang. 2019. The Impact of Media Censorship: 1984 or Brave New World? *American Economic Review* 109, 6 (2019).
[10] Justin Clark, Robert Faris, and Rebekah Heacock Jones. 2017. Analyzing Accessibility of Wikipedia Projects Around the World. *Berkman Klein Center Research Publication* 2017-4 (2017).
[11] Ronald Deibert, John Palfrey, Rafal Rohozinski, Jonathan Zittrain, and Janice Gross Stein. 2008. *Access Denied: The Practice and Policy of Global Internet Filtering*. MIT Press, Cambridge.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[13] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.
[14] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.
[15] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.
[16] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 609–614.
[17] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1489–1501.
[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751.
[19] Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Review* (2013), 326–343.

[20] Gary King, Jennifer Pan, and Margaret E Roberts. 2017. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review* 111, 3 (2017), 484–501.

[21] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 138–143.

[22] Rebecca MacKinnon. 2012. *Consent of the Networked: The Worldwide Struggle For Internet Freedom.* Basic Books, New York.

[23] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 615–621.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[25] Evgeny Morozov. 2011. *The Net Delusion: The Dark Side of Internet Freedom.* PublicAffairs, New York.

[26] Jason Ng. 2013. Who's the Boss? The difficulties of identifying censorship in an environment with distributed oversight A large-scale comparison of Wikipedia China with Hudong and Baidu Baike. *Citizen Lab* (2013). https://citizenlab.ca/2013/08/a-large-scale-comparison-of-wikipedia-china-with-hudong-and-baidu-baike/

[27] Daniel Oberhaus. 2017. Wikipedia's Switch to HTTPS Has Successfully Fought Government Censorship. *Motherboard* (2017). https://bit.ly/2T5aEWm

[28] Jennifer Pan and Margaret E. Roberts. 2019. Censorship's Effect on Incidental Exposure to Information: Evidence from Wikipedia. *SAGE Open* (2019).

[29] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 446–457.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[31] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 529–535.

[32] Margaret E. Roberts. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall.* Princeton University Press, Princeton.

[33] Emma Rodman. 2019. A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis* (2019), 1–25.

[34] Sergey Sanovich, Denis Stukal, and Joshua A Tucker. 2018. Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia. *Comparative Politics* 50, 3 (2018), 435–482.

[35] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *stat* 1050 (2017), 22.

[36] Arthur Spirling and P Rodriguez. 2019. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. (2019).

[37] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019). https://doi.org/10.18653/v1/p19-1159

[38] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

[39] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10–29.

[40] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 53–59.

[41] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. IEEE, 1521–1528.

[42] Shih-Ming Wang and Lun-Wei Ku. 2016. ANTUSD: A large Chinese sentiment dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2697–2702.

[43] Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. NEZHA: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204* (2019).

[44] Victoria Baranetsky Welinder, Yana and Brandon Black. 2015. Securing Access to Wiki-media Sites with HTTPS. *Wikimedia Blog* (2015). https://diff.wikimedia.org/2015/06/12/securing-wikimedia-sites-with-https/

[45] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[46] Jane Zhang. 2019. How Baidu built an encyclopedia with 16 times more Chinese entries than Wikipedia. *South China Morning Post* (2019). https://www.scmp.com/tech/big-tech/article/3038402/how-baidu-baike-has-faced-against-wikipedia-build-worlds-largest

[47] Xiaodong Zhang and Mark Boukes. 2019. How China's flagship news program frames "the West" : Foreign news coverage of CCTV's Xinwen Lianbo before and during Xi Jinping's presidency. *Chinese Journal of Communication* 12, 4 (2019), 414–430.

[48] Xiaoquan Michael Zhang and Feng Zhu. 2011. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101, 4 (2011), 1601–15.

[49] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 15–20.

A ADDITIONAL SENTIMENT ANALYSIS RESULTS

A.1 Model Accuracy on Validation Set

In training the TextCNN models, we held out 20% of our training set as a validation set. The validation set was used to assess the quality of the models during training. The model with the best accuracy on the validation set in each run was selected as the outputted model. A.1 reports the average accuracy (over 10 runs) of the models on the validation sets.

Table A.1: Model Accuracy on Validation Sets

2-class		
	Baidu Baike	90.29
	Wikipedia	89.65
	People's Daily	92.64
3-class		
	Baidu Baike	67.44
	Wikipedia	66.07
	People's Daily	67.80

Note: "2-class" classification means that the training and validation sets contain only negative and positive headlines. "3-class" classification additionally has neutral headlines included.

A.2 Sentiment Analysis Results with Neutral Headlines Included

Table A.2: Model Accuracy

	Model	Accuracy
Naive Bayes		
	Baidu Baike	56.42
	Wikipedia	55.63
	People's Daily	57.79
SVM		
	Baidu Baike	55.53
	Wikipedia	55.29
	People's Daily	54.71
TextCNN		
	Baidu Baike	61.71
	Wikipedia	60.89
	People's Daily	58.55

Table A.3: Wikipedia vs. Baidu Baike

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.11	0.00	-0.06	0.00	-0.03	0.12
Democracy	-0.08	0.00	-0.04	0.04	-0.02	0.23
Election	-0.09	0.00	0.00	0.87	-0.01	0.62
Collective Action	-0.10	0.00	-0.06	0.00	0.00	0.89
Negative Figures	-0.05	0.00	-0.01	0.47	0.03	0.02
Social Control	0.01	0.59	0.03	0.08	0.03	0.04
Surveillance	-0.06	0.00	-0.05	0.00	0.01	0.51
CCP	0.05	0.00	0.03	0.01	0.04	0.01
Historical Events	-0.04	0.02	-0.01	0.66	0.02	0.05
Positive Figures	0.08	0.00	0.07	0.00	0.08	0.00

Table A.4: Wikipedia vs. People's Daily

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.17	0.00	-0.07	0.00	-0.05	0.01
Democracy	-0.13	0.00	-0.07	0.00	-0.06	0.00
Election	-0.13	0.00	0.00	0.93	-0.01	0.53
Collective Action	-0.15	0.00	-0.06	0.00	-0.02	0.22
Negative Figures	-0.02	0.17	0.00	0.96	0.01	0.32
Social Control	0.05	0.00	0.02	0.22	0.00	0.97
Surveillance	-0.01	0.61	-0.04	0.02	-0.01	0.56
CCP	0.04	0.01	0.04	0.00	0.03	0.02
Historical Events	-0.01	0.53	0.00	0.78	0.03	0.00
Positive Figures	0.10	0.00	0.06	0.00	0.10	0.00

A.3 Sentiment Analysis Results Comparing Baidu Baike and People's Daily

A.5 reports the results comparing models trained on Baidu Baike and those trained on People's Daily, where Baidu Baike is model *a* and People's Daily is model *b*. A positive coefficient means that on average People's Daily model rates a given category more positively than Baidu Baike.

A.6 reports results from the same comparison but with headlines with neutral labels included in the training and test sets.

Table A.5: Baidu Baike vs. People's Daily (2-class)

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.09	0.00	-0.02	0.48	-0.07	0.00
Democracy	-0.05	0.05	-0.01	0.68	-0.02	0.29
Election	-0.03	0.31	0.04	0.08	-0.02	0.36
Collective Action	-0.06	0.01	0.02	0.28	-0.01	0.57
Negative Figures	0.05	0.02	0.01	0.69	-0.04	0.04
Social Control	0.03	0.09	0.01	0.72	-0.02	0.27
Surveillance	-0.03	0.25	-0.02	0.49	-0.02	0.24
CCP	0.04	0.04	0.00	0.82	-0.01	0.33
Historical Events	0.04	0.07	0.01	0.46	0.01	0.72
Positive Figures	0.07	0.00	-0.01	0.35	0.00	0.92

Table A.6: Baidu Baike vs. People's Daily (3-class)

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.07	0.00	-0.01	0.64	-0.02	0.21
Democracy	-0.06	0.01	-0.03	0.17	-0.04	0.04
Election	-0.04	0.07	0.00	0.93	0.00	0.88
Collective Action	-0.06	0.00	0.00	0.84	-0.02	0.26
Negative Figures	0.03	0.07	0.01	0.44	-0.02	0.20
Social Control	0.04	0.02	-0.01	0.59	-0.03	0.04
Surveillance	0.05	0.00	0.01	0.55	-0.02	0.20
CCP	-0.01	0.63	0.01	0.46	0.00	0.73
Historical Events	0.03	0.06	0.00	0.88	0.01	0.36
Positive Figures	0.02	0.01	-0.01	0.34	0.01	0.12

B FURTHER DETAILS ON THE TNEWS DATASET

The TNEWS Dataset comprises of 73,360 Chinese news headlines from Toutiao, a Chinese news and information content platform. The dataset contains news headlines from 15 categories: *story, culture, entertainment, sports, finance, house, car, education, technology, military, travel, world, stock, agriculture* and *gaming*.

The TNEWS dataset is part of the Chinese Language Understanding Evaluation (CLUE) Benchmark, which serves as a common repository of datasets used to test the accuracy of trained models. (For an equivalent of CLUE in English, see GLUE: https://gluebenchmark.com/). Because the length of a news headline is usually short, the TNEWS dataset is widely used as either training or testing data for machine learning models that tackle short-text classification tasks. Given that the downstream task we are interested in is the classification of news headlines, the TNEWS dataset serves as the ideal source of data in our case.

The TNEWS dataset is split into a training set (53,360 headlines), a validation set (10,000 headlines) and a test set (10,100 headlines). For our purpose, we pooled the three sets and randomly selected 5,000 news headlines from the pooled set. Because the news headlines are not labeled according to sentiment in the dataset, we manually labeled the sentiment of the headlines in our selected subset. Each headline is labeled by two independent coders of native Chinese speaker and any conflict in labeling is resolved.

C LIST OF TARGET WORDS

Freedom (自由) = {自由 (freedom), 言论自由 (freedom of speech), 集会自由 (freedom of assembly), 新闻自由 (freedom of the press), 结社自由 (freedom of association), 自由权 (right to freedom), 民主自由 (democracy and freedom), 自由言论 (free speech), 创作自由 (creative freedom), 婚姻自主 (marital autonomy), 自由民主 (freedom and democracy), 自由市场 (free market), 自决 (self-determination), 自决权 (right to self-determination), 生而自由 (born free), 自由自在 (free), 自由选择 (freedom of choice), 自由思想 (freedom of thought), 公民自由 (civil liberties), 自由竞争 (free competition), 宗教自由 (freedom of religion), 自由价格 (free price)}

Election (选举) = {选举 (election), 直接选举 (direct election), 议会选举 (parliamentary election), 间接选举 (indirect election), 直选 (direct election), 换届选举 (general election), 民选 (democratically elected), 投票选举 (voting), 全民公决 (referendum), 总统大选 (presidential election), 大选 (election), 普选 (universal suffrage), 全民投票 (referendum), 民主选举 (democratic election)}

Democracy (民主) = {民主 (democracy), 自由民主 (freedom and democracy), 民主自由 (democracy and freedom), 民主制度 (democratic system), 民主化 (democratization), 社会民主主义 (social democracy), 民主运动 (democratic movement), 民主主义 (democracy) , 民主改革 (democratic reform), 民主制 (democratic system), 民主选举 (demoratic election), 民主权力 (democratic rights), 多党制 (multi-party system), 民主法制 (democracy and rule of law), 民主权利 (democratic rights)}

Social Control (维稳) = {维稳 (social control), 处突 (emergency handling), 社会治安 (public security), 反恐怖 (counter-terrorism), 公安工作 (police work), 预防犯罪 (crime prevention), 收容审查 (arrest and investigation), 治安工作 (public security work), 大排查 (inspections), 扫黄打非 (combating pornography and illegal publications), 接访 (petition reception), 反邪教 (anti-cult)}

Surveillance (监控) = {监控 (surveillance), 监测 (monitor), 监视 (surveillance), 管控 (control), 监看 (monitor), 监视系统 (surveillance system), 截听 (tapping), 监控中心 (surveillance center), 情报服务 (intelligence service), 排查 (inspection), 监视器 (surveillance equipment), 情报搜集 (intelligence collection), 间谍卫星 (reconnaissance satellite) , 管理网络 (internet control), 监控器 (surveillance equipment), 监控站 (surveillance center), 监控室 (surveillance center), 数据采集 (data collection)}

Collective Action (抗议) = {抗议 (protest), 示威 (demonstration), 示威游行 (demonstration; march), 示威抗议 (demonstration; protest), 游行示威 (demonstration; march), 静坐示威 (sit-in), 绝食抗议 (hunger strike), 请愿 (petition), 示威运动 (demonstration), 游行 (demonstration; march), 罢教 (strike), 静坐 (sit-in), 集会游行 (demonstration; assembly), 罢课 (strike), 签名运动 (signature campaign)}

Positive Figures (党和国家) = {毛泽东 (Mao Zedong), 江泽民 (Jiang Zemin), 胡锦涛 (Ju Jintao), 习近平 (Xi Jinping), 周恩来 (Zhou Enlai), 朱[illegible]基 (Zhu Rongji), 温家宝 (Wen Jiabao), 李克强 (Li Keqiang), 邓小平 (Deng Xiaoping), 曾庆红 (Zeng Qinghong), 华国锋 (Hua Guofeng), 李鹏 (Li Peng), 杨尚昆 (Yang Shangkun), 谷牧 (Gu Mu), 吴邦国 (Wu Bangguo), 李岚清 (Li Lanqing), 纪登奎 (Ji Dengkui), 乔石 (Qiao Shi), 邹家华 (Zou Jiahua), 李瑞环 (Li Ruihuan), 俞正声 (Yu Zhengsheng), 张高丽 (Zhang Haoli), 田纪云 (Tian Jiyun), 回良玉 (Hui Liangyu), 李源潮 (Li Yuanchao), 贾庆林 (Jia Qinglin), 姚依林 (Yao Yilin), 张立昌 (Zhang Lichang), 尉健行 (Wei Jianxing), 姜春云 (Jiang Chunyun), 李铁映 (Li Tieying), 王兆国 (Wang Zhaoguo), 罗干 (Luo Gan), 刘靖基 (Liu Jingji), 杨汝岱 (Yang Rudai), 王光英 (Wang Guangying), 彭佩云 (Peng Peiyun), 刘云山 (Liu Yunshan), 丁关根 (Ding Guangen), 彭真 (Peng Zhen), 胡启立 (Hu Qili), 曾培炎 (Zeng Peiyan), 何东昌 (He Dongchang)}

Negative Figures = {林彪 (Lin Biao), 王洪文 (Wang Hongwen), 张春桥 (Zhang Chunqiao), 江青 (Jiang Qing), 姚文元 (Yao Wenyuan), 刘晓波 (Liu Xiaobo), 丹增嘉措 (Tenzin Gyatso), 李洪志 (Li

Hongzhi), 陈水扁 (Chen Shui-bian), 黄之锋 (Joshua Wong), 黎智英 (Jimmy Lai), 艾未未 (Ai Weiwei), 李登辉 (Lee Teng-hui), 李柱铭 (Martin Lee), 何俊仁 (Albert Ho), 陈方安生 (Anson Chan), 达赖 (Dalai Lama), 陈光诚 (Chen Guangcheng), 滕彪 (Teng Biao), 魏京生 (Wei Jingsheng), 鲍彤 (Bao Tong)}

CCP (中国共产党) = {党中央 (central committee), 中国共产党 (CCP), 党支部 (party branch), 中共中央 (central committee), 共青团 (CCP youth league), 共青团中央 (youth league central committee), 党委 (party committee), 中央党校 (central party school)}

Historical Events = {抗日战争 (Anti-Japanese War), 解放战争 (China's War of Liberation), 抗美援朝 (the War to resist U.S. Aggression and Aid Korea), 改革开放 (Reform and Opening up), 香港回归 (Hong Kong reunification), 长征 (Long March), 三大战役 (Three Great Battles in the Second Civil War), 秋收起义 (Autumn Harvest Uprising), 南昌起义 (Nanchang Uprising), 澳门回归 (Transfer of sovereignty over Macau), 志愿军 (Volunteer Army), 土地改革 (Land Reform), 六四 (June Fourth Movement), 遵义会议 (Zunyi Conference), 九二南巡 (Deng's Southern Tour in 1992), 广州起义 (Guangzhou Uprising), 西藏和平解放 (Annexation of Tibet), 井冈山会师 (Jinggangshan Huishi), 百团大战 (Hundred Regiments Offensive), 文革 (Cultural Revolution), 文化大革命 (Cultural Revolution), 大跃进 (Great Leap Forward), 四人帮 (Gang of Four), 解放农奴 (Serfs Emancipation)}

D LISTS OF PROPAGANDA ATTRIBUTE WORDS

Positive Adjectives = {稳定, 繁荣, 富强, 平稳, 幸福, 振兴, 发展, 兴旺, 昌盛, 强盛, 稳当, 安定, 局势稳定, 安定团结, 长治久安, 安居乐业}

Negative Adjectives = {动荡, 衰落, 震荡, 贫瘠, 不幸, 衰退, 萧条, 败落, 没落, 衰败, 摇摆, 不稳, 时局动荡, 颠沛流离, 动荡不安, 民不聊生}

E EXAMPLES OF EVALUATIVE ATTRIBUTE WORDS

Positive Evaluative = {情投意合, 精选, 严格遵守, 最根本, 确有必要, 重镇, 直接接管, 收获, 思想性, 均需参加, 可用于, 当你落后, 同意接受, 居冠, 感化, 完美演出, 急欲, 多元地理环境, 形影不离的朋友, 一举击败, ...}

Negative Evaluative = {金融波动, 科以, 畸型, 向.. 开枪, 破碎家庭, 撬动, 头皮发麻, 颠覆, 迟疑, 血淋淋地, 驱赶, 干的好事, 责骂不休, 生硬, 㓥蚀, 拉回, 走失的家畜, 燃眉之急, 喷溅, 违反, ...}

For the full list of evaluative words from the augmented NTU sentiment dictionary (ANTUSD), see https://academiasinicanlplab.github.io/#resources.