Classification of Chinese State-Generated Sockpuppet Tweets

April 09, 2022

Jacob Barkow
jacobbarkow@berkeley.edu

Anna Cheng
anna.cheng@berkeley.edu

Abstract

Platforms such as Twitter have seen increased foreign state-linked sockpuppets influencing online discourse. This project trained classification models on sockpuppet tweet data to identify when a tweet originates from a bad-faith information operation run by the Chinese state media. Results show that high classification accuracy (F1 = 0.97) can be achieved through the use of BERTweet sentence embeddings. Future improvements can be made with access to more labelled data and experimentation with multilingual support.

1 Introduction

In recent years, Twitter and other social media platforms have been the target of public criticism due to infiltration of inauthentic sockpuppet accounts linked to foreign information operations aimed at spreading propaganda. [1] Sockpuppet accounts differ from normal state-linked Twitter accounts (such as communication or media officials) because they are disingenuous in identity, often posing as regular netizens, and the propagandizing opinions they spread can manipulate the political discourse of Twitter and US foreign policy writ large.

To fight these efforts, the Twitter Transparency Center has released 37 datasets of purged inauthentic accounts attributed to manipulation campaigns from 17 countries, spanning more than 200 million Tweets, to provide insight on their efforts to moderate such content. One such country, the People's Republic of China (PRC), should be especially scrutinized, as China has long been "suspected of hiring as many as 2,000,000 people [known as "Wumaos"] to surreptitiously insert huge numbers of pseudonymous and other deceptive writings" into social media sites. [2] As modern US-China relations grow increasingly tense, with raging online debates on topics ranging from the pro-democracy protests in Hong Kong in 2018, the Xinjiang - Uyghur conflict, and COVID-19, CCP-backed sockpuppeting is only increasing. [4]

As these operations continue to expand, the motivation for being able to identify when a tweet or user is part of such an operation is clear: to keep an open, safe, and healthy social media ecosystem safe from manipulation. To pursue this goal, we leveraged a pre-trained BERTweet language model to train neural networks on the classification task of recognizing when a tweet originates by a CCP-backed sockpuppet account vs. a legitimate good faith actor/user.

2 Related Work

While there is prior work on sockpuppet classification as a machine learning task, existing research does not match our specific goal (identifying political actors), platform (Twitter), and methods (NLP), perhaps due to the lack of available labeled datasets. Solorio et. al. (2013) [5] assembled a classifier 239 features that captured the presence of linguistic characteristics to detect whether a given Wikipedia article was written by a sockpuppet, achieving an F1 score of 0.72. Bu et. al [6] used link analysis between accounts in multiple Chinese and English forums to build a classifier of suspected sockpuppet accounts; the highest accuracy achieved was 0.39.

For CCP sockpuppeting, there has been both qualitative and quantitative of the problem, though no attempts to build a classifier specifically for identiying CCP sockpuppets. The Stanford Internet Observatory analyzed the Twitter dataset we use in our report, finding that the June 2020 dataset focused heavily on topics of the Hong Kong pro-democracy protests, COVID, and exiled billionaire Guo Wengui. [7] Bolsover and Howard (2017) [8] scraped Twitter and Chinese social media site Weibo and found evidence of automated anti-Chinese posts on Twitter, presumably to convince Chinese users of overseas bias against China. Zervopoulos et al. (2020) [9] used NLP methods for fake news detection related to the Hong Kong protests, scoring an F1 of 0.92 - but they look at the veracity of news posts as opposed to whether a tweet originated from a sock-puppet account. Given the wide range and diversity of model results, problem types, and social media platforms examined, model performance on our specific task is not easy to benchmark.

3 Dataset

Our project faced a dilemma in that our primary data source, the Twitter Information Operations dataset on identified state-backed accounts, contained observations of sockpuppet accounts only (i.e. the positive class, which we assigned as $isCCP = 1$). We thus constructed the full data set for classification by identifying and building examples for the negative class.

We hypothesized that there would be differences in model performance based on the construction of the dataset. Taking into consideration which negative examples to include, we collated two datasets, each with a different negative class (which we call $AllChina$ vs. $ProChina$). Our two datasets are built from the following data:

isCCP (1) Subset of the Twitter Information Operations source; filtering on English tweets only, matching keywords for Chinese politics and current events; from March 2019 and later

AllChina (0) Random sample of all tweets on China across Twitter matching the criterion used to filter on the Twitter Info Ops dataset

ProChina (0) Tweets from a specific network of users (US leftists, political activists, Chinese diaspora), who are ideologically more likely to express pro-China/CCP opinions

3.1 Positive Class: isCCP

The *isCCP* dataset is a collection of tweets from Chinese sockpuppets from the Twitter Information Operations hub, which we combined into a complete dataset of 14,263,532 tweets. All individual datasets are listed in Table 1.

After downloading the raw files from Twitter, we subsetted the tweets based off the following criteria:

- English language (as BERTweet only handles English) - tthis filter caused the most dropout, as only 17 percent of the data were in English

Release Date	Focus
August 2019	HK Protests
September 2019	General
June 2020	General
December 2021	Xinjiang/Changyu

Table 1: Raw data with links to Twitter documentation

- Matching keywords on specific Chinese current events, foreign policy, etc. i.e. Hong Kong, Xinjiang, COVID, Guo Wengui, as outlined by the Stanford Internet Observatory report

- Posted in March 2019 or later (i.e. the beginning of the Hong Kong protests) - some accounts were posting non-political content as early as 2008, and were likely acquired and repurposed by the CCP in recent years

After subsetting, we were left with **19,367 tweets**, which is the number of tweets we also mine for the negative class.

3.2 Negative Class: AllChina

We created our first set of negative examples on a dataset of tweets containing discourse on China from all of Twitter (*AllChina*), with a specific focus on political and current event discourse during the documented time of CCP sockpuppet activity. To assemble the negative examples, we scraped Twitter using the searchtweets API wrapper to find tweets that matched the isCCP filter criteria and then randomly sampled from the query a subset of 19,367 tweets. The tweets do not overlap with isCCP as the CCP tweets were purged from Twitter and are no longer viewable or accessible via API. There may be some tweets that appear in this sample that Twitter has not identified as state-backed, but we assume the proportion is very small, relying on Twitter's reported sockpuppet purge rates.

This main drawback of the dataset is we hypothesized it to be more negative about China on average, as Twitter and other US-based social media platforms have been observed to have had a very strong Sinophobic and often anti-Communist bent. [9] [10] Further, Chinese mainland citizens—who are most likely to express positive sentiment about China and the CCP—are blocked from being able to post on Twitter. In order to ensure we do not create a biased classifier with high accuracy but poor precision (i.e. any "positive" tweet on China is immediately classified as a sockpuppet bot), we also built a ‘ProChina‘ dataset to experiment on, with the expectation that a more tailored dataset would express more similar sentiment to the CCP sockpuppet tweets.

3.3 Negative Class: ProChina

To create a dataset that more closely resembles the ideologies expressed in the isCCP dataset, i.e. more positive or sympathetic towards China, we leveraged graph theory by seeding a network with names of users who are more likely to express pro-CCP or “pro-China” opinions[1] and crawled within this network to find the top 500 node accounts in terms of edges - similar to the “friend graph” algorithms used by Facebook. The seed nodes users were a mix of the following communities:

- Chinese diaspora activists, with a focus on anti-colonization, anti-imperialism, anti-Sinophobia activism
- American Communists: Marxist-Leninist-Maoists, Communists, and other proponents of authoritarian Socialism; “tankies”
- Leftist media outlets (Jacobin, The Intercept)

This resulted in a list of the Top 500 most followed accounts in these communities, which we considered as representative of the most popular beliefs in their communities. Not all users surfaced by this method are fully pro-CCP or pro-China (naming the dataset *ProChina* is shorthand), but ideologically these individuals are more likely to support the CCP and/or be skeptical of mainstream Western perspectives on China. In fact, some accounts that surface from our network analysis are actually Chinese state-affiliated media officials (not considered sockpuppets because they disclose their state connections) and/or are users who have been retweeted or quote tweeted in our isCCP dataset, offering strong evidence for the validity of our method.

We then scraped tweets from the top 500 accounts matching the same filters as before, and again sampled 19,367 tweets from this corpus.

3.4 Final Datasets

From our 3 classes of data, we create 2 final datasets: 1) ‘ProChina‘ and 2) ‘AllChina‘, both with the same ‘isCCP‘ tweets as the positive class, but with differing negative class samples. The total number of observations for each dataset is 38,734 tweets.

3.4.1 Dataset Comparability

Before we trained any models, we sought to justify our selection of the datasets, ideally by establishing some sort of a quantitative measure for classifying tweets as pro-China, or anti-China. The thorough approach would be to create and train another “pre”-classifier to first classify each set of tweets as pro- or anti- China or CCP to ensure that the tweets are balanced in sentiment. Due to limitations in our resources, we instead attempted two naive methods - 1) topic-based sentiment analysis, and 2) annotation of a random sample of our two datasets. Overall, our analysis showed that on average the ProChina dataset did contain more more pro-CCP and pro-China discourse than the overall AllChina dataset. A more complete analysis can be found in **Appendix A**.

4 Models

4.1 BERTweet Tokenization and Embeddings

To represent the tweet data numerically, we fed the corpus into BERTweet to generate vector representations of each tweet. BERTweet is a BERT variant developed by Nguyen et al. (2020) in their paper BERTweet: A pre-trained language model for English Tweets. BERTweet was trained on a corpus of 845 million English tweets and BERTweet embeddings have been shown to perform better than their RoBERTa counterparts with tweet-specific NLP tasks. We opted to make use of the pooled output embedding in which each tweet is represented as a 768-length vector. We did not fine-tune the BERTweet embeddings due to 1) the smaller size of our dataset, and 2) the extremely strong model results generated using pre-trained embeddings. Future iterations of this research would benefit from pursuing fine-tuning.

Prior to generating the embeddings, we normalized and tokenized the tweets so the inputs could be understood by the BERTweet model. We used the same Twitter-specific tokenizer and normalization as employed in the original paper; this tokenizer was specifically tailored for tweet data

[1] We group ”Pro-China” and ”Pro-CCP” attitudes together in our work; the CCP has a high approval rate among its citizens (tracked by the Harvard Ash Center) and is in many ways synonymous with the Chinese people and country. Similarly, anti-China and anti-CCP content is also closely bucketed - blatant Sinophobia is still a blemish on the CCP’s project of global dominance.

Dataset	F1	Recall	Precision
ProChina	0.9431	0.9431	0.9436
AllChina	0.9379	0.9379	0.9385

Table 2: Baseline results

and cleanly handles idiosyncrasies such as user @-tagging, retweets, emoji, and internet slang.

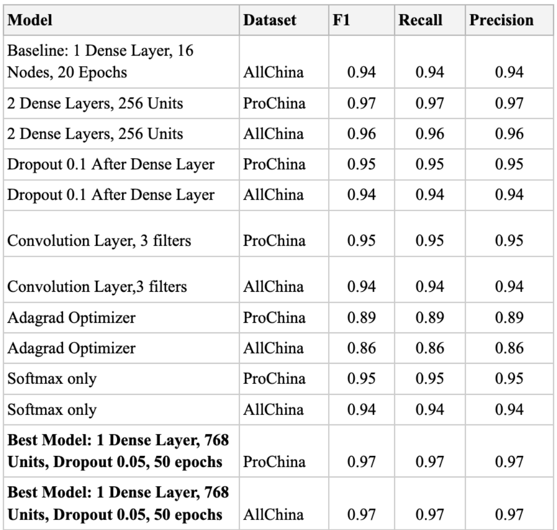
4.2 Baseline

Using the BERTweet embeddings, we trained a single-layer neural network to produce a baseline model. This initial layer contained 256 nodes with a relu activation, and the initial model employed a standard Adam optimizer. From this architecture, we obtained the F1 scores of 0.94 for the ProChina dataset and the AllTwitter dataset when testing against our test set, with closely similar Precision and Recall values, as seen in **Table 2**.

4.3 Additional Experiments

To improve upon our baseline results, we explored different numbers and node size of dense layers, adding convolutional layers, and employing more aggressive dropout. We also tested replacing the Adam optimizer with the Adagrad optimizer. Overall, we did not see much variation in performance with our hyperparameter tuning process.

We were able to achieve a maximum F1 score of 0.97 on both the Pro-China and All-China datasets with the following model. Key model results are shown in the following image, with the last row containing the best model results:



Model	Dataset	F1	Recall	Precision
Baseline: 1 Dense Layer, 16 Nodes, 20 Epochs	AllChina	0.94	0.94	0.94
2 Dense Layers, 256 Units	ProChina	0.97	0.97	0.97
2 Dense Layers, 256 Units	AllChina	0.96	0.96	0.96
Dropout 0.1 After Dense Layer	ProChina	0.95	0.95	0.95
Dropout 0.1 After Dense Layer	AllChina	0.94	0.94	0.94
Convolution Layer, 3 filters	ProChina	0.95	0.95	0.95
Convolution Layer,3 filters	AllChina	0.94	0.94	0.94
Adagrad Optimizer	ProChina	0.89	0.89	0.89
Adagrad Optimizer	AllChina	0.86	0.86	0.86
Softmax only	ProChina	0.95	0.95	0.95
Softmax only	AllChina	0.94	0.94	0.94
Best Model: 1 Dense Layer, 768 Units, Dropout 0.05, 50 epochs	ProChina	0.97	0.97	0.97
Best Model: 1 Dense Layer, 768 Units, Dropout 0.05, 50 epochs	AllChina	0.97	0.97	0.97

5 Results

Even without BERTweet fine-tuning or training models with other architectures, both models on both datasets performed extremely well, averaging around an F1 score of 0.95. We observed that hyperparameter tuning had little effect on model performance. Furthermore, training the same model parameters on different datasets did not result in different results as we initially hypothesized: the sets of F1 scores for both datasets were always within mere significant figures of another. Interestingly, when testing the ProChina model on the AllChina dataset, the F1 score for class 0 (i.e., not CCP) dropped to 0.90, but the reverse did not experience show any performance drop. This finding suggests that despite the fact that “overall” Twitter attitudes to be more mixed towards China, the model trained on general data did not necessarily make more errors classifying real pro-China users as sockpuppets.

To further investigate the reason for such high F1 scores, we removed all specialized neural net architecture and tried a single basic soft-max layer, which still generated F1 scores of 0.94 and 0.95 for the ProChina and AllChina datasets respectively. As performance remained high without advanced model architecture, possible reasons for the performance could be attributed to any of (or combination of) the following: 1) the strength of the pre-trained BERTweet embeddings; 2) the potentially trivial nature of this specific task; and lastly, 3) issues in the construction of the dataset itself.

For example, limiting our analysis to the English language forced us to drop out more than 80 percent of the total tweets in the original dataset. That model performance remains almost identical on both ProChina and AllChina leads us to suspect that we may have introduced bias in our dataset by inadvertently building a model that picks up on linguistic habits between non-native English speakers vs. native English speakers (especially since we did not fine-tune BERTweet on our specific dataset that does contain more ESL speakers than on average), transforming a complicated task to a trivial one. Another consideration is that it may have been more representative of Twitter’s true environment to create a dataset with class imbalances (after all, there would not be an even distribution of sockpuppets and real users on Twitter), and then apply class-balancing techniques such as SMOTE on a dataset that we have designed as unbalanced.

5.1 Error Analysis

As part of our results analysis, we also looked at the types of classification errors of our highest performing model. Because the model is a neural net, we could not ascribe causality or meaning to any single feature, but qualitative analysis could be performed to understand where the model may be lacking. We found that our model made errors common to NLP classification tasks, such as detection of negation and tone (sarcasm); more uniquely, the model also seemed to mix up tweets of "soft power" (i.e. appreciation of Chinese culture, history, or natural landscapes) and general news tweets mentioning China as false positives. A full analysis can be found in **Appendix B**.

6 Conclusion

In our project, we architected two datasets to run model experiments on, and ultimately built a strong classifier with an F1 score of 0.97, observing little variable impact of the datasets on model performance. We explored many reasons for why model performance could be so high, and overall caution that the score seems to be an overestimate of if we were to run the classifier in the true Twitter environment.

The number of future possibilities of further research show that political sockpuppet detection is a complex task – especially when considering the multilingual nature of foreign actors. Additional work could take on a number of possible avenues, such as: looking into training a model on Chinese tweets as well and comparing models trained on the multilingual $BERTbase_{Chinese}$ embeddings with our results trained on BERTweet embeddings to see what the differences between languages are (as well as examining tradeoffs, considering that the multilingual models don't use a Twitter corpus); finetuning BERTweet embeddings and continuing to limit the analysis to English only; attempting other model architectures such as SVMs and Random Forest; experimenting with dataset architecting by introducing intentional class imbalance; additional feature engineering of non-linguistic content such as metadata or network links; and more.

References

[1] A. Bessi and E. Ferrara. 2016. "Social bots distort the 2016 US presidential election online discussion." First Monday, Volume 21, Number 11 - 7 November 2016

[2] Twitter Safety Blog. 2021. "Disclosing state-linked information operations we've removed"

[3] G. King, J. Pan, and M. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." In American Political Science Review 111(3): 484–501

[4] S. Myers, P. Mozur, and J. Kao. 2022. "Bots and Fake Accounts Push China's Vision of Winter Olympic Wonderland."

[5] T. Solorio, R. Hasan, and M. Mizan. 2013. "A Case Study of Sockpuppet Detection in Wikipedia." In Proceedings of the Workshop on Language Analysis in Social Media, 59–68. Association for Computational Linguistics.

[6] Bu, Zhan, Xia, Zhengyou, Wang, Jiandong. 2013. "A sock puppet detection algorithm on virtual spaces." In Knowledge-Based Systems. 37. 366-77. 10.1016/j.knosys.2012.08.016.

[7] Stanford Internet Observatory. 2020. "Sockpuppets Spin COVID Yarns: An Analysis of PRC-Attributed June 2020 Twitter takedown."

[8] A. Zervopoulos, A. G. Alvanou, K. Bezas, A. Papamichail, M. Maragoudakis, and K. Kermanidis. 2020. "Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter." In: Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, 584.

[9] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, and R. Tromble. 2020. "Detecting East Asian Prejudice on Social Media." In Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 162–172, Online. Association for Computational Linguistics.

[10] F. Tahmasbi , L. Schild , C. Ling, J. Blackburn , G. Stringhini , Y. Zhang, and S. Zannettou. 2021. Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19.

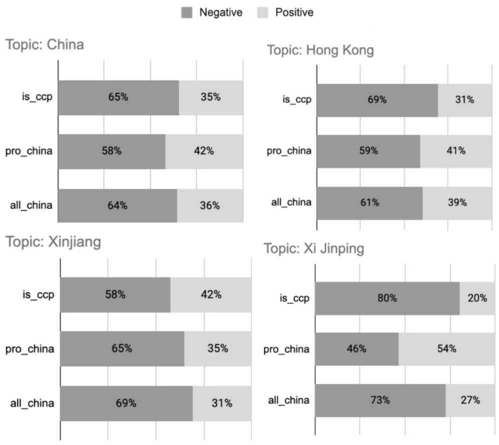
[11] S. Bhatia, 2018. "Topic-Specific Sentiment Analysis Can Help Identify Political Ideology." EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis.

Appendix

A Sentiment Analysis for Dataset Comparability

A.1 VADER

Bhatia et al (2018) [11] find political ideology can be represented by a characteristic sentiment distribution over different topics. With this methodology, we leveraged the pre-trained VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Intensity Analyzer classifier in the NLTK package, which is uniquely trained on social media posts and is able to handle emojis/text emoticons, internet slang/abbreviations and other social media quirks, to see whether the distribution of sentiment on certain topics in each dataset was comparable. A sampling of topic cluster sentiment results are shown below.

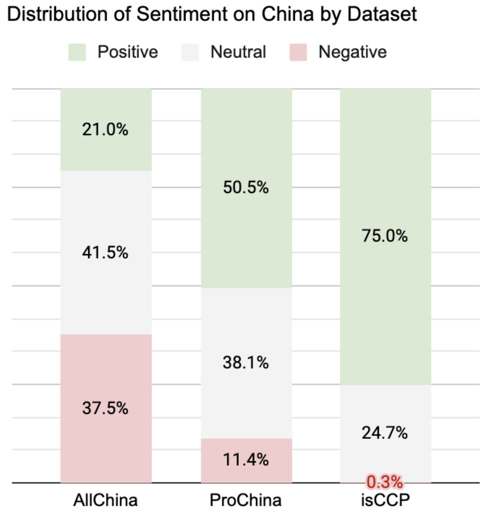


As seen above, the results are not only inconsistent, but also nonsensical, as the CCP should be espousing favorable views of both China and Xi Jinping, but per VADER, the majority of CCP tweets expressed negative sentiment. It becomes clear that this classifier is naive, for the following reasons: 1) tweet sentiment is not necessarily a 1:1 translation of political belief: for example, a pro-CCP tweet might be expressing outrage at "biased treatment" of China in the political arena, and as the tweet expresses negative emotions, it would be classified as containing negative sentiment, but ultimately be a positive view on China itself; and 2) this may be exacerbated by the well known phenomena that users tend to be more negative on social media, as outrage tends to receive more engagement.

A.2 Classification by Hand-Label

As the VADER sentiment classifier method did not generate clear results, we then sampled 1,000 tweets from each data set (approximately 5 percent of the 60,000 total unique tweets scraped) and annotated them by hand. This method also generated more granularity, as we were able to classify tweets as Positive, Negative, or Neutral. Tweets were classified as neutral if they discussed Chinese culture, such as media, entertainment, or food; expressed a "fence-sitting opinion" (i.e. "China is both good and bad"); was a tweet sharing a news article related to China but not expressing any opinion; or if it was otherwise irrelevant to politics or current events.

The results are displayed in the figure below:



With this sampling method, we confirmed that our ProChina dataset expresses more support for China and the CCP compared to the general population of Twitter users; the neutral category contained many tweets that discussed news about China, but also had far more "fencesitting" opinions that seemed conflicted. Unsurprisingly, the CCP dataset is majority positive towards China, with approximately a third of the tweets being neutral and virtually zero expressing negative sentiment toward China. Finally, the general Twitter userbase seems to be more evenly distributed in the sentiment categories for their stance on China, but in their negative category contained more extreme tweets veering into blatant racism, xenophobia, and anti-Communist sentiment.

B Error Analysis

Errors can be broadly grouped under the following categories, with example tweets:

B.1 Errors common among False Negatives

Negation A relatively common issue in NLP, this category consists of tweets that often contains language that comes off as critical of China and the CCP, but ultimately rejects such claims, as in the following false negative tweet:

> “The claim of “since 2017, the Chinese government has rounded up over one million Uyghurs and other Turkic and Muslim people in detention camps in the Xinjiang region”is from dubious source...”

The strength of criticism in the embedded quote in question may have thrown off the classifier, even though the goal of the tweet is to ultimately reject the claim.

Tone Another example of an error category were errors of tone, such as sarcasm:

> “According to the Western mainstream media, China always has nefarious motives behind its schemes...”

Out of context, this tweet seems extremely negative towards China; however, the author of this tweet is actually criticizing Western media outlets (and their portrayal of China) as opposed to citing them to establish credibility. This context is missed, and the model is confident that it is not written by the CCP sockpuppets.

Fencesitting The classifier also seemed to struggle with tweets that participated in fence-sitting, i.e. tweets that did not firmly establish a “side”:

> “@StateDept America faces many of the same problems that China faces today. There are constant fiscal crisises and financial crisises.”

B.2 Errors common among False Positives

Soft power The model seemed confused with the number of CCP tweets that aimed to influence through ”soft power”, including sharing educating aspects of Chinese culture and history (such as festivals, traditions, proverbs/sayings), praising the beauty of the Chinese country landscapes and vacation destinations, etc, and made some errors classifying them as non-CCP, as with the following:

> ”Get Lost in a Snow-Blanketed Wonderland this Friday with @chinaorgcn’s Stunning Photo Essay on Frosty Scenery at the Tianshan Mountains in Xinjiang...”

News News updates posted by the CCP were also common among the mistakes made by the model, even news with the obvious intention of showcasing efficacy or goodwill of the CCP government.

> ”China’s average utilization rate of hydropower, wind and photovoltaic power had achieved 97 percent, 96 percent, and 98 percent respectively. . . ”