Using Stylometry to Assess Chinese Influence on English-language Sources

Michelle Gong

Abstract

Fake news and propaganda are a prevalent issue on websites and messaging apps popular with overseas Chinese diaspora. These sites and apps are integral parts of life for this demographic, as well as the only source of news for many of them.

In this paper, I study the linguistic features of Chinese state-sponsored articles as they compare to Western sources on the same topic. Three topics of a controversial nature were selected: Hong Kong, Taiwan, and Xinjiang. Data was collected from tweets, headlines, and article bodies related to these topics, and this data is then analysed for linguistic features such as word usage and frequency. I also propose a method to train a neural network to apply the analysis to a long-form article of unknown authorship to detect a level of Chinese state media influence.

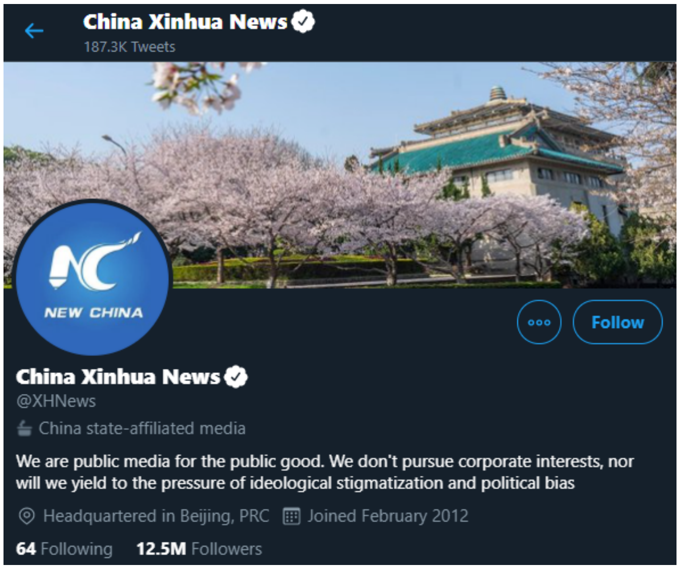
The results of the above evaluation showed that there are different word usages between China-sanctioned Xinhua and overseas media like NPR, and there are certain words, phrases, and omissions that could be used to detect Chinese media influence on an article of unknown origin.

Introduction

A prevalent issue among the mainland Chinese diaspora is the spread of sensationalist “fake news” and propaganda through various websites and popular messaging apps, many of which are essential parts of life and the only source of news for older Chinese immigrants (Funke 2018). Though fake news is hardly a problem limited to the mainland Chinese diaspora, this particular situation is unique due to the strict censorship laws placed on media by the ruling Chinese Communist Party (”CCP”). More than a dozen government bodies review media and set guidelines restricting what political sensitive topics may be covered and how, so as to hide calls for reform or exposure of human rights abuses that may subvert CCP authority (Xu and Albert 2017).

For the reasons listed above, there is a general distrust among the diaspora of state-sanctioned media such as *Xinhua* or *People’s Daily*, but this same suspicion does not extend to all Chinese-language news, and even less to English-language sources. This is where the problem lies, as English-language versions of major CCP media exist and even have a presence in spaces long-banned to Chinese citizens.

In this work, I seek to formulate an approach that can detect possible Chinese media influence in unbranded (ie. not connected to any large or known news source) English-language articles. By adapting methods used to detect American political polarity in long-form text (Saligrama 2019), the goal is to create a ”spectrum of influence” on which news articles could be placed, according to how similar they seem in bias and viewpoint to state-sanctioned Chinese media. Through this, the hope is that the effects of CCP propaganda on overseas diaspora can be reduced or, at the very



China Xinhua News
@XHNews
China state-affiliated media

We are public media for the public good. We don't pursue corporate interests, nor will we yield to the pressure of ideological stigmatization and political bias

Headquartered in Beijing, PRC Joined February 2012

64 Following 12.5M Followers

Figure 1: Profile for the Xinhua News Agency on Twitter. Note the ”state-affiliated” warning flag put on the account by the site. This account’s main audience would be overseas viewers, as Twitter is banned in China.

least, there can be increased awareness of the sort of rhetoric used to try to influence opinions on key issues.

Background

An initial survey of articles on the English-language version of *Xinhua News* revealed (to an untrained, human eye) what appears to be a general unified writing style that is the result of the CCP's strict guidelines. This style is particularly apparent when contrasted with Western sources such as NPR or the BBC, and is glaringly obvious when the topics involved are sensitive and controversial. The following provides a brief overview of some key issues that are relevant and used throughout the course of this work.

Xinjiang is a northwestern region of China home to about 11 million Uyghurs, one of China's 55 officially recognised ethnic minorities. Uyghurs are predominately Muslim and speak a Turkic language with their own modern script separate from Chinese *hanzi* script. Since 2017, in an attempt to enforce homogenisation of language and culture under the Han Chinese majority, the CCP have subjected Uyghurs in Xinjiang to surveillance, religious and linguistic restrictions, ”reeducation camps”, and various other human rights abuses that have led to rebuke and sanctions by members of the international community (Maizland 2020).

Taiwan is an island off the southern coast of China that has been governed independently from mainland China since 1949. The citizens of Taiwan enjoy such freedoms as an open internet and free elections, and were not subject to the One-Child Policy in effect until recently. Mainland China views Taiwan as part of it under the ”One China Principle”, or ”one country, two systems” framework. Though there has always been friction due to this, the election of pro-independence Taiwanese president Tsai Ing-wen brought issues to the forefront, and tensions have been rising to the point of Beijing threatening force should Taiwan try to assert independence (Albert 2020).

Hong Kong is a special administrative region of China that, like Taiwan, enjoys political, economic, and personal freedoms not afforded those in mainland China. Ceded to Great Britain in 1842 after China's defeat in the First Opium War, Hong Kong was returned in 1997 with an agreement that the region would continue to enjoy its freedoms for 50 years under ”one country, two systems”. Despite it being far yet from the 50-year mark, Beijing has in recent years attempted to assert control over Hong Kong. This led to mass protests in 2014 and 2019, and unrest from its national security law of 2020 is still ongoing (Albert and Maizland 2020).

In looking at news related to the above topics, some patterns appear. *Xinhua*, for example, regularly in Taiwan-related articles makes heavy emphasis on words like “secessionist” and “provocative”, and employs scare quotes around the words “Taiwan independence”. News about Xinjiang (if any) have no mention of detention camps or human rights abuses. Meanwhile, NPR articles on similar topics highlight the tension between Beijing and Taiwan, and the former's willingness to use force if necessary, and Xinjiang articles are written completely based around the human rights abuses being carried out. Hong Kong news are on *Xinhua*'s side about ”rioters” while on NPR ”pro-democracy”– the list goes on, but the fact remains that this is all, once again, purely from a human perspective, with all its innate biases. The goal of this paper is to compare Chinese and Western reporting of various sensitive topics to see if there are any major linguistic features separating them, and to use these linguistic features to explore a method of evaluating Chinese media influence.

The contributions of this paper are: 1. I create an objective method of analysing the form and influence of CCP's state-sanctioned news. 2. I apply existing research on AI detection of authorship and writing styles (stylometry). 3. I collect articles, tweets, headlines, and various other writing samples from Chinese and Western media on the above mentioned topics, and use that data to perform analyses on word usage and style.

Related Work

Stylometry

Stylometry is the application of the study of linguistic style that can be used to determine authorship of unknown documents (Juola 2008). Various tests exist and are used with the help of the Natural Language Toolkit, a platform that allows the creation of Python programs to work with human language data.

Mendenhall's Characteristic Curves of Composition looks simply at individual word *length*, under the belief that individuals will have a ”characteristic curve” of word length usage that would become precise and constant over the course of their lifetime.

Kilgariff's Chi-Squared Method measures “distance” between words, analysing word and word combination usage of a known corpus and unknown text. The two corpora are merged into a single corpus, then the *n* most common words are found. Then these common words are used to calculate how often they might have been expected to appear in each of the two original corpora if they had been from the same author, and then a chi-square distance is calculated.

John Burrows's Delta method also measures “distance”, but compares an unknown text with many different authorships rather than just one. It creates a large corpus of any number of authors' works and then measures how a given anonymous text and this large corpus of texts diverge from the average of all of them put together. This method gives equal weight to every feature it measures to avoid having common words disproportionately affect results, as is oft the issue with chi-square tests.

Applications

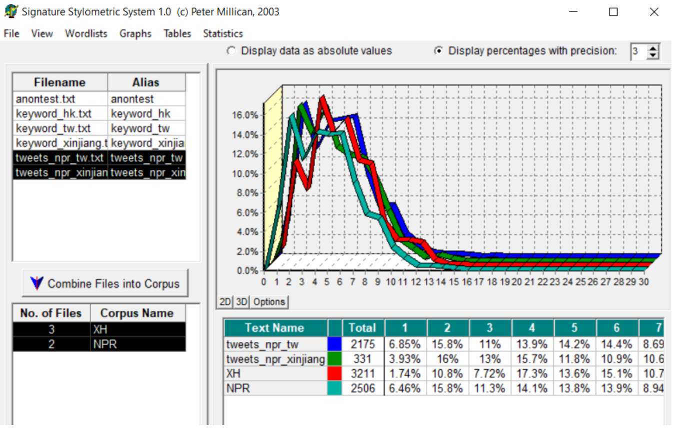
Some of the most common applications have been in determining authorship of individual pieces of the *Federalist Papers*, such as one method that utilised the relative frequencies of a set of three works (Fung 2003), and another that made use of neural networks (Tweedie, Singh, and Holmes 1996). The latter combined statistical methods with neural networks to produce a powerful classification tool.

In the Chinese-language sphere, (Ng, Feldman, and Peng

2020) used linguistic data from Chinese social media site Weibo (similar to Twitter) to predict what posts will be marked for censorship. In accordance with China's censorship laws, any posts on Weibo that might subvert CCP authority are often censored and taken down. In their research, Ng *et al.* also built a neural network classifier that predicted censorship, and found a set of linguistics features that play a role in censorship. Among the examples for posts that were censored include phrases such as (translated) ”spirit of democracy”, which as will later be shown aligns with some of the findings in English-language sources.

In the closest to what I am trying to accomplish, Know-Bias is a training and inference method for detecting political bias in long-form text. Tweets were collected en masse and annotated to serve as training data for a two-step classification scheme: neutral sentences are removed from articles, then the remaining text is put to a polarity classifier to determine political bias (Saligrama 2019).

There also exist various programs designed to facilitate stylometric analyses. Signature is freeware designed by Dr. Peter Millican of Oxford University that evaluates wordlists, various chi-square analysis, among other facilities (Millican 2014).



Filename	Alias
anontest.txt	anontest
keyword_hk.txt	keyword_hk
keyword_tw.txt	keyword_tw
keyword_xinjiang.t	keyword_xinjia
tweets_npr_tw.txt	tweets_npr_tw
tweets_npr_xinjian	tweets_npr_xin

No. of Files	Corpus Name
3	XH
2	NPR

Text Name		Total	1	2	3	4	5	6	7
tweets_npr_tw	🟦	2175	6.85%	15.8%	11%	13.9%	14.2%	14.4%	8.69
tweets_npr_xinjiang	🟩	331	3.93%	16%	13%	15.7%	11.8%	10.9%	10.6
XH	🟥	3211	1.74%	10.8%	7.72%	17.3%	13.6%	15.1%	10.7
NPR	🟦	2506	6.46%	15.8%	11.3%	14.1%	13.8%	13.9%	8.94

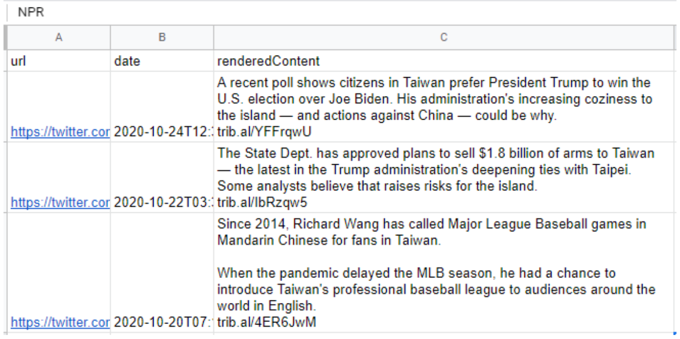
Figure 2: Example of how Signature displays and analyses texts

Problem Statement

This paper seeks to compare linguistic features of CCP-sanctioned English-language media and Western media on the same topics. In order to keep the data to a manageable amount, the sources this time were limited to two news outlets: *Xinhua* and NPR (National Public Radio) news. *Xinhua* is China's official, state-run press, while NPR is perhaps the closest the United States has to a ”official”, government-sponsored network à la the BBC. These two sources will serve as two extremes of Chinese influence and propaganda.

The data used will mostly be in the form of tweets. NPR tweets headlines from its main site almost one-to-one, and scraping tweets is a much more straightforward process than scraping off the main NPR website. *Xinhua* data will also come from its official Twitter account, but given how its website is designed, it is much more conducive to web scraping, and so headlines and whole articles where able will also be collected.

All the above data will be collected by me through the use of web scraping program ParseHub as well as a snscrape script for tweets.



NPR

A	B	C
url	date	renderedContent
https://twitter.cor	2020-10-24T12::	A recent poll shows citizens in Taiwan prefer President Trump to win the U.S. election over Joe Biden. His administration's increasing coziness to the island — and actions against China — could be why. trib.al/YFFrqwU
https://twitter.cor	2020-10-22T03::	The State Dept. has approved plans to sell $1.8 billion of arms to Taiwan — the latest in the Trump administration's deepening ties with Taipei. Some analysts believe that raises risks for the island. trib.al/IbRzqw5
https://twitter.cor	2020-10-20T07::	Since 2014, Richard Wang has called Major League Baseball games in Mandarin Chinese for fans in Taiwan. When the pandemic delayed the MLB season, he had a chance to introduce Taiwan's professional baseball league to audiences around the world in English. trib.al/4ER6JwM

Figure 3: Sample from one of the datasets used. This is a collection of tweets from NPR in which the word ”Taiwan” appears.

Problem Solution

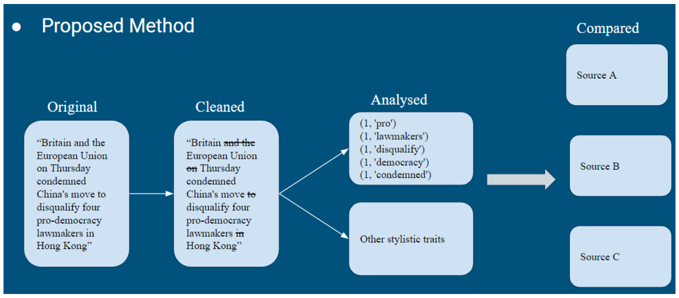
Data was collected and separated according to source and topic (ex: ”NPR, Hong Kong”, or ”Xinhua, headline, Taiwan”). Using snscraper scripts, tweet archives for Xinhua and NPR were searched for key words ”hong kong”, ”taiwan”, and ”xinjiang”, all tweets with the searched words were collected into two files: an Excel spreadsheet with source URL, content, and data for archival purposes, and a .txt file with only the tweet content for actual processing. Xinhua's news site was also scraped for about 20 pages of headlines on each topic.

Each of these data subsets were then run through a text analyser that output the most common words, bigrams (two word combinations), and trigrams (three word combinations), while cleaning up, or ignoring, common ”stopwords”, or filler words such as ”this” and ”but”. This output data was then cleaned up manually as needed. For example, in the cases of tweets, attached URLs were not always parsed out cleanly, and so any presented results affected by such URLs were removed.

In total, about 6,000 to 7,000 tweets were collected from *Xinhua*, and a little under 1,000 collected from NPR on the various topics. The top ten to twelve most frequent words, bigrams, and trigrams were taken from each of the three topics, and in the case of tweets a direct comparison was made between Xinhua and NPR's word and phrase usage.

Conclusion

Through analysis of the datasets, I was able to find that there does exist certain linguistic features unique to Xinhua and NPR respectively. For example, on the topic of Hong Kong, Xinhua consistently and repeatedly refers to the region as the



- Proposed Method

Original

“Britain and the European Union on Thursday condemned China's move to disqualify four pro-democracy lawmakers in Hong Kong”

Cleaned

“Britain ~~and the~~ European Union ~~on~~ Thursday condemned China's move ~~to~~ disqualify four pro-democracy lawmakers ~~in~~ Hong Kong”

Analysed

(1, 'pro')
(1, 'lawmakers')
(1, 'disqualify')
(1, 'democracy')
(1, 'condemned')

Other stylistic traits

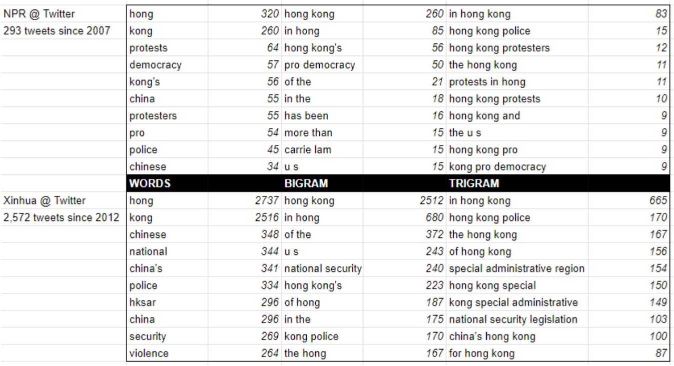
Compared

Source A

Source B

Source C

Figure 4: Proposed method for analysing and comparing articles



NPR @ Twitter	hong	320	hong kong	260	in hong kong	83
293 tweets since 2007	kong	260	in hong	85	hong kong police	15
	protests	64	hong kong's	56	hong kong protesters	12
	democracy	57	pro democracy	50	the hong kong	11
	kong's	56	of the	21	protests in hong	11
	china	55	in the	18	hong kong protests	10
	protesters	55	has been	16	hong kong and	9
	pro	54	more than	15	the u s	9
	police	45	carrie lam	15	hong kong pro	9
	chinese	34	u s	15	kong pro democracy	9
	WORDS		**BIGRAM**		**TRIGRAM**	
Xinhua @ Twitter	hong	2737	hong kong	2512	in hong kong	665
2,572 tweets since 2012	kong	2516	in hong	680	hong kong police	170
	chinese	348	of the	372	the hong kong	167
	national	344	u s	243	of hong kong	156
	china's	341	national security	240	special administrative region	154
	police	334	hong kong's	223	hong kong special	150
	hksar	296	of hong	187	kong special administrative	149
	china	296	in the	175	national security legislation	103
	security	269	kong police	170	china's hong kong	100
	violence	264	the hong	167	for hong kong	87

Figure 5: A comparison of word usage on the topic of Hong Kong between NPR and Xinhua’s respective Twitters.

”Hong Kong Special Administrative Region”, the specific phrase appearing around 200 times in 2,500 tweets. Meanwhile, in NPR’s 293 tweets on Hong Kong, the phrase does not even make the top ten most frequent words or phrases; for the most part, the region is simply, ”Hong Kong”. Similarly, ”democracy” does not make an appearance in Xinhua’s ten most frequent, but ”democracy” is mentioned 57 times in NPR’s 293 tweets. Similar patterns were found in the other topics as well, but limited comparison on the part of NPR (only 10 tweets for Xinjiang and 115 for Taiwan) make true analysis difficult.

What Worked

A great amount of data was able to be extracted and collected, and I learned a lot about web and tweet scraping, and preparing data for processing. Through these and the analyses of word frequency, I was able to find some key stylistic patterns and at least affirm some of what I originally suspected.

What Did Not Work

Due to lack of personal experience and ability, the scope of this project was limited to analysing word and phrase frequency. I was not able to apply any real neural network or AI concepts to create something that could analyse any unknown article.

Obstacles and Limits

As should be expected given that it is China’s official news source, Xinhua’s tweets and headlines on the various topics far outnumbered those available in NPR. At lowest, NPR’s archive only contained 10 tweets related to Xinjiang, most of which relate to the recent detention camps in the region, making analysis skewed and difficult. In addition, Twitter’s API limits the amount of tweets scraped to roughly 2000.

Another limit to keep in mind is the fact that in processing the data, all punctuation was stripped out. As mentioned in the introduction, one linguistic feature of *Xinhua* articles is its use of scare quotes around words such as Taiwanese independence. The elimination of this in the processing removes an interesting and key feature of how *Xinhua* articles are styled.

Perhaps the greatest limit this project faced was the use of preexisting stylometry programs and libraries due to my own limited abilities to create them from scratch. In this way, I could not customise them for this specific purpose, and that limited my scope and very likely created skewed results as well.

Future Work

There is great potential for future work on this topic. Due to my own limited abilities, I was not able to create a neural network classifier that might be able to further take and analyse the collected data. However, in addition to creating this, there is great potential for a merge of research on AI, machine learning, and international relations. One possible avenue to take this research would be to expand the datasets to include many more news sources from all around the world, such as the United Kingdom’s BBC, France’s France24, and Russia’s TASS. Given China’s growing global presence, its relationships with various countries are also changing, for better or worse, and it would be interesting to see how this might affect public news outlets’ reporting on sensitive issues. From a glance at France24, for example, there is an almost belligerent, combative tone towards China’s human rights abuses, the wording much stronger than any NPR article. This seems to go along with the fact that China and French relations have been increasingly sour as of late. I would like to further explore this line of research.

References

Albert, E., and Maizland, L. 2020. Council on Foreign Relations: Democracy in Hong Kong. https://www.cfr.org/backgrounder/democracy-hong-kong. [Online; accessed 12-December-2020].

Albert, E. 2020. Council on Foreign Relations: China-Taiwan Relations. https://www.cfr.org/backgrounder/china-taiwan-relations. [Online; accessed 12-December-2020].

Fung, G. 2003. The disputed federalist papers: Svm feature selection via concave minimization. New York, NY, USA: Association for Computing Machinery.

Funke, D. 2018. On WeChat, rogue fact-checkers are tackling the app’s fake news problem. https://www.poynter.org/fact-checking/2018/on-wechat-rogue-fact-checkers-are-tackling-the-apps-fake-news-problem. [Online; accessed 12-December-2020].

Juola, P. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval* 1:233–334.

Maizland, L. 2020. Council on Foreign Relations: China's Repression of Uighurs in Xinjiang. https://www.cfr.org/backgrounder/chinas-repression-uighurs-xinjiang. [Online; accessed 12-December-2020].

Millican, P. 2014. The Signature Stylometric System. http://www.philocomp.net/texts/signature.htm. [Online; accessed 12-December-2020].

Ng, K. Y.; Feldman, A.; and Peng, J. 2020. Linguistic fingerprints of internet censorship: the case of sinaweibo.

Saligrama, A. 2019. Knowbias: Detecting political polarity in long text content. *CoRR* abs/1909.12230.

Tweedie, F.; Singh, S.; and Holmes, D. I. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities* 30:1–10.

Xu, B., and Albert, E. 2017. Council on Foreign Relations: Media Censorship in China. https://www.cfr.org/backgrounder/media-censorship-china. [Online; accessed 12-December-2020].