Digital Humanities - Experiment with method needed for final project

Evaluating the South China Morning Post with Deep Learning + Natural Language Processing

Jasen Lo

*Hello reader, this project is easier to read in a Google Colab Notebook:
https://colab.research.google.com/drive/1xkpP0azVmy_WdN93q1fDe7VyNRM2dsUN?usp=sharing

Introduction

The South China Morning Post (SCMP), regarded as the most credible English-language newspaper in Hong Kong, was acquired by Chinese multinational technology company, Alibaba Group, in 2016. Due to the close relationship between the Chinese government and Alibaba Group, critics have since accused the newspaper of losing its editorial independence and neutrality, instead prioritising the promotion of Chinese soft-power due to its new ownership.

> *My research question is focused around questions examining how SCMP has changed in editorial style or reportage coverage since Alibaba Group acquired SCMP in 2016.*

Methodology

South China Morning Post (SCMP) positions itself as a politically neutral newspaper, and is commonly used as an authoritative source of information in English on affairs in Hong Kong and Greater China by the English speaking communities all over Asia, as well as foreign observers and media organisations outside of Asia.

The experiment with method presented in this notebook is an initial attempt to use BERT, a deep learning NLP model, to find out if the SCMP has a establishment-friendly or establishment hostile journalistic slant on aggregate. I decided to use other media publications with clear journalistic slants such as China Daily and Hong Kong Free Press (HKFP) as proxies for pro-establishment or anti-establishment sentiments respectively. China Daily is an English-language daily newspaper owned by the Publicity Department of the Chinese Communist Party while HKFP is a free, non-profit news website specifically founded to to cover Hong Kong's pro-democracy movement and to counteract Hong Kong's deteriorating press freedom.

Thus, this methodology experiment aims to find out if using BERT is a viable way to evaluate the SCMP's overall political slant. The high level steps I took to conduct this experiment:

1. **Web scraping** - I scraped 22,000 SCMP articles from 2019, 10,000 articles from China Daily (2018 - 2020), 15,000 articles from the HKFP (2018 - 2020).
2. **Training BERT** - Training BERT to classify China Daily vs. Hong Kong Free Press articles (pro vs. anti-establishment).
3. **Evaluating Model** - Evaluate the trained BERT model is good at identifying HKFP and China Daily articles using a validation set.
4. **Classify SCMP Data** - Use the trained BERT model to see the distribution of pro vs. anti establishment sentiments in SCMP articles.

Data Biography

Web Scrap Method

The SCMP, HKFP and ChinaDaily datasets were built through web scraping methods.

For SCMP and HKFP, I used a sitemap library to walk through the domains' sitemaps in order to find all of the URLs of all the articles on the respective websites. I then used Request, BeautifulSoup4 libraries to scrap each URL that was found. On each request, I scraped the article title, author, date published, summary, and in the case of SCMP, topics as well.

For China Daily, I was able to make use of the structure of their website that loaded more articles by clicking a "Load More" Button. I used the Selenium library to simulate a real person on a browser and clicked the button until no more articles were found. Then, I downloaded the HTML code of the fully-expanded page. I then scraped all of article title, author, date published, and summary data from this HTML page.

All of the code is available in the repository here.

Dataset Summaries Two datasets were used in this method experimentation:

1. HKFP + China Daily Dataset - This dataset serves as the training, testing, and validation set that trains BERT to recognise articles that are either pro-establishment (China Daily) or anti-establishment (HKFP). Dataset has 4000 rows × 2 columns. 4000 entries of 2000 rows from each publication and 2 columns:
    1. Title - Titles of the China Daily or HKFP articles
    2. Class - Labels to identify if article is HKFP or China Daily. **1 is for HKFP, 0 is for China Daily.**
2. SCMP Dataset - This dataset serves as the dataset of interest whose data I feed into BERT for classification. Dataset has 2453 rows × 2 columns. 2453 entries from the SCMP and 2 columns:
    1. Title - Titles of the SCMP article
    2. Summary - Summary of the SCMP article (for looking into classification nuances, if necessary)

Data Curation Notes

1. SCMP has a total of 1.4 million article urls.Decisions had to be made about how to choose the relevant articles that could be evaluated due to time and computational resource constraints. I thus only scraped URLs containing the sub-string 'hong-kong', as well as articles from 2019 (chosen due to the more politically fraught media environement).
2. For HKFP and China Daily, which had approximately 100,000 and 10,000 articles of interest. I thus limited the data down to 2,000 articles from 2018 to 2020 for both datasets. I have yet to decide on a representative and *accurate* process of dataset construction, these measures are mostly for the purposes of saving time and computational resources for the time being. Only articles directly relevant to Hong Kong news was chosen for these datasets. Note that the combination of these two datasets create the training, testing, and validation set used to fine-tune BERT.

Evaluation of Methodology

Using deep learning in this context might not be such a great idea, more ideas need to be tested to decide on viability. Perhaps a more 'distant reading'focused approach would suit this research question better?...

The fine-tuned BERT model that is trained on the HKFP + China Daily dataset overwhelmingly classifies SCMP articles to have HKFP or anti-establishment sentiment (2447:6). This surprinsing result can be explained by several reasons, raising conerns about this implementation, the dataset construction, or the methodological assumptions in general.

1. SCMP articles really are more similar to HKFP articles. However, this does not necessitate that SCMP and HKFP have similar journalistic slants, as this similarity can be rooted in non-political factors such as style of writing, formatting and naming conventions, and other such natural language nuances. This reason raises concerns about whether or not the methodological assumption that HKFP and China Daily can serve as proxies for pro and anti-establishment journalistic slants is a valid assumption.
    - Moving Forward: This assumption could viably be tested with the complete training set, as the results might have been partially skewed by the small training set size.
2. Only SCMP articles with 'Hong Kong' in their urls were included in this experiment. However, this excludes those articles in which the term 'Hong Kong' is not in the title of the article, but is nonetheless pertaining to Hong Kong. This is an important conern since HKFP and SCMP are based in Hong Kong and writes about Hong Kong from a local perspective, while China Daily is based in Beijing and frequently writes about Hong Kong from a distanced perspective, which may have contributed to the model's confidence that most SCMP articles were like HKFP articles.
    - Moving Forward: As far as I'm aware, there are not English-language pro-establishment publications based in Hong Kong. However, this concern might be slightly balanced by including articles from foreign publications in the anti-establishment datasets should they reliably publish anti-establishment articles.

Unfortunately, early returns from this methodology are not too promising. It might be more fruitful to use other NLP methods to answer the research question instead of relying on deep learning methods. The unclean SCMP datatset contains topics, for example, which could be suited to a data-viz heavy 'distant reading' approach analysis.

LO Appendix

#dataconstruction - I constructed the two datasets used in this method exploration using web scraping methods. I also cleaned these datasets to remove duplicates and missing entries. However, the real challanage of dataset construction was thinking through which articles should be included to reflect the methodological assumptions that the training set articles reflect their corresponding journalistic slants and what sorts of articles should the target dataset be made of. Additional decisions had to be made about sampling from this larger dataset and thus reducing the size of the dataset, while attempting to make sure that the data in the training set for both classes were comparable. I honestly don't think I effectively applied this LO given the lackluster performance of the trained model, which can be largely explained by a less than ideal data construction process, which could have benefitted from much larger dataset sizes and training times. However, it is difficult to do so while continously adjusting what data should be chosen and evaluating each iteration of each dataset construction.

#toolsandtechniques - Used web scrapping, data cleaning, natural language processing and deep learning libraries. For web scrapping, I abidded by web scraping principles to declare the user agent as well as sleeping the scraper such as not to overload the website's servers. I was also able to effectively apply BERT to carry out the classification that I was atttempting to do from pre-processing data into tokens to training to classfying. Having an understanding of BERT helped! However, the returning insights of this classfication were hampered by poor data construction and potentially invalid methodological assumptions.

#dhquestion - This exploration did not tackle the original research question, as it was more of a viability check on whether or not the method was effective. However, it is important to note that this viability check is connected to the original research question, except carried out at a much smaller time/data scale. Still, these findings provide additional decision making impetus as to how the research question could potentially be answered with other methods instead, or perhaps an improved iteration of the current implementaion.

Feedback Questions For Prof. Ostrow

1. I think I'm not sure if I should pivot to more "Underwood"-like approach and give up the deep learning approach. The results I got were somewhat discouraging, as they showed how much trial-and-error I might have to do with #dataconstruction in putting a dataset together that can actually answer the original research question. Even then, I'm not entirely sure that this methodology, even if I put together the perfect dataset, will work. At the same time, I can see how a 'distant reading' approach will be equally, if not more, time-

consuming. Given what you've seen in this exploration and your understanding of what 'distant-reading' entails, what are your thoughts on which approach might be worth investing more time into?

2. Something that bothered me throughout this exercise was how my model might be overfitting, and how confident the model was at classifying anything I threw at it. It is this lack of uncertainty, that irks my statistical training, as I imagine that if I were a reader of my future work, I would remain skeptical of these overly-confident classifications. Yet, I also note how 'distant reading' avoids these arriving at quantitative conclusions because of the method is much more descriptive than explanatory. How can I think about balancing the worries of expressing self-doubt in my findings, yet being able to arrive at something more than just a descriptive analysis?

Techincal Stuff Here

↳ 41 cells hidden