https://doi.org/10.2352/issn.2168-3204.2023.20.1.6

Data Extraction, Visualization, and Storytelling: A Case Study in Headline Analysis on "The Hongkong News" with Deep Learning

Michael Kin-Fu YIP; The Chinese University of Hong Kong Library; Hong Kong
Vincent Wai-Yip LUM; The Chinese University of Hong Kong Library; Hong Kong

Abstract

This paper presents the methodologies to extract the headline and illustrations from a historical newspaper for storytelling to support digital scholarship. It explored the ways in which new digital tools can facilitate the understanding of the newspaper content in the setting of time and space, "The Hongkong News" was selected from Hong Kong Early Tabloid Newspaper for the case study owing to its uniqueness in historical value towards the scholars. The proposed methodologies were evaluated in OCR (Optical Character Recognition) with scraping and Deep Learning Object Detection models. Two visualization products were developed to showcase the feasibility of our proposed methods to serve the storytelling purpose.

Introduction

Digital scholarship tools nowadays give scholars multiple options to explore the complex past. They can browse through an interactive timeline, study the massive digitized textual content or explore some of the large events via the newspaper headlines.

This project explores issues of narrative time and space. It embraces a conception of geography, of space, as the place can discover the contemporary lives that were linked up to the newspaper articles and time is another fundamental axis of narrative which can be employed for storytelling for digital scholarship. It explores the ways in which new digital tools allow us to understand the newspaper content in the setting of time and space, ways for geography to meet history, via some sophisticated text-mining computational algorithms. The idea is to enable scholars to access the content from various facets, with a view to understanding the articles in the newspaper in a sort of non-conventional approach.

The "Hong Kong Early Tabloid Newspapers"《香港早期小報》, launched in 2022, has collected tabloid newspapers published in Hong Kong during the 20th century. Contrary to the serious broadsheet newspapers, the tabloid newspapers represent the leisure and entertainment of the masses, covering topics like politics, operas, dramas, comics, and pornographies etc. Here, *The Hongkong News* from our *Hong Kong Early Tabloid Newspapers Collection* was analysed through the above-mentioned approaches. The Japanese Occupation holding of The Hongkong News started publication in 1941, right after the Christmas Day surrender of the British Crown Colony, and lasted up to August 17, 1945, the very last week before Hong Kong's liberation. The newspaper had previously launched in 1940 at separate premises in the Crown colony. *The Hongkong News* as published during the Japanese Occupation offers scholars the voice of Japan from Hong Kong. The uniqueness of this newspaper was also discussed in [1] which showed case the important historical value of it towards the scholars.

Through extracting headlines from *The Hongkong News,* two visualization products were developed. Using this project as a case study, we will share the experience that we encountered while working through these newspaper headlines. Our contributions are as follows:

- Extracting visual and textual content from historical newspapers.
- Applying deep learning algorithms in headline extraction.
- Developing two visualization products that facilitate public and scholar access to our existing repository.

Motivation

The Hongkong News contained a wealth of information. Though scholars were interested in the content of this newspaper, they were currently restricted to manual search over each page. For instance, when users wanted to search headlines that mentioned specific locations, they needed to browse over thousands of pages in our repository. Besides, current digital assets could be further utilized to provide value apart from preservation and online access. There were lots of space to optimize the usage of these digitized images. Beyond the image, our team will apply various computational techniques to explore more value in this tabloid newspaper. Another challenge for the users is to gain concrete insight from the newspaper content if the corpus concerned is at a large scale. We try to address this challenge by performing headline analysis which would be a useful tool for newspaper topic analytics as echoed by [2].

To further promote tabloid newspapers' multifaceted content and improve its accessibility, there were two main objectives in our project. First, two semi-automatic procedures were provided to recognize and extract the newspaper headline. Though various tools in data extraction have existed, an evaluation of these strategies was meaningful for the community to understand how to choose these tools. Comparing two headline extraction strategies, our project stated their varieties in cost and technique. Second, two visualization products were developed to facilitate users accessing our digital repository. Our visualization products not only give scholars insight to headline that appear at different times and spaces but also enable them to address humanities questions, such as "Which battle was the Japanese propaganda focus on at a specific time?" and "How was the Japanese propaganda changed during the war?".

This project covered 530 issues in *The Hongkong News* from the year 1942 to 1945. The related dataset was available at CUHK Research Data Repository [3]. The digital image of *The Hongkong News* was available in our Digital Repository [4]. The visualization

product can be available at our Digital Scholarship Project [5]. The sample source code is also available in GitHub with links accessible in [5].

Related Work

Headline Analysis

Thanks to digitization and the OCR work, some in-depth textual analysis can be made feasible. For example, in [6], we have performed similar textual analysis in a tabloid newspaper through the Name Entity Recognition (NER) and presented some sort of visualizations in Word Cloud and Geospatial temporal presentations. Likewise, most of the studies in newspaper analysis focus on textual analysis at the article level. Typically Headline Analysis in newspaper is one of such tools. For instance, [7-12] have used headline analysis as a qualitative analysis for the full news story in newspapers. The headline is found to be an important part of a news story and its summary to catch the reader's attention. The analysis of the headlines has produced complementary and convergent findings with the corpus analysis. In one of the studies [2], the author concluded that analyzing headlines was proven to be a good "down-sampling" option to reduce large news corpora to a manageable amount of data. But most of the studies were still using conventional methods and when the newspaper corpus was at a large scale, it would be an issue here. Automatic or semi-automatic means of headline extraction will turn out to be necessary.

Object Extraction

Conventionally, given the varying layout and style of printed materials, paper headlines could only be extracted manually. Without computational support, the workflow was time-consuming and labour-intensive. Though some commercial OCR software could help to recognize text, researchers still need to search and retrieve headlines from OCR results. Headlines still could not be extracted automatically.

Due to the use of Convolutional Neural Network (CNN), lots of improvements have been made in object detection in recent years. Many object detectors have been built to address the user demand. Object detection has been used by [13] and [14] to automatically detect different illustrations in modern publications.

In 2020, Library of Congress applied Faster-RCNN model to extract information from newspaper pages in their broad-scale project [15]. Their training dataset contains up to 3,600 pages with 48,409 annotations. The model has been trained for 17 hours on the NVIDIA T4 GPU. Though the precision of headline extraction was near 75% in average precision in validation set, this model could not generally fit other data. It had a relatively low generalization ability. Given that the visual content recognition model has been trained on World War 1-era newspapers, the model was repurposed to 19th-century newspapers, a dropoff in the performance of headline recognition resulted (AP:21.2% for 1850-1875 newspapers and AP:51.6% for 1875-1900 newspapers). It is difficult to repurpose the model to other corpus given the non-generic nature of the dataset. Added to this, Detectron2 model was used in their studies, we are using YOLO recognition model here to see if the performance can be further improved.

Headline and Image Extraction

Two strategies have been considered in this section. By comparing two methods: (1) OCR software with scraping, and (2) the Self-developed YOLOv5 model, we aim at demonstrating the feasibility of headline extraction in different contexts.

(1) OCR software with scraping

In strategy one, we mainly adopted OCR (Optical Character Recognition) and web scraping. OCR is a commonly used technology in recognizing text within the digital image. ABBYY FineReader is one of the most popular commercial OCR software. Though ABBYY FineReader has high accuracy in text recognition, it did not provide any function for extracting specific content. In this section, our team will explore the way of using ABBYY FineReader to transform digital images to HTML format, and using Beautiful Soup, a python scraping library, to pull headlines from HTML. In conclusion, there were two steps:

- Using ABBYY FineReader to extract HTML.
- Scraping through Beautiful Soup.

By using ABBYY FineReader, all digital images were converted to HTML format at the first step. In the extracted HTML, all information was mixed. Therefore, further data processing is necessary to distinguish headline content from other information.

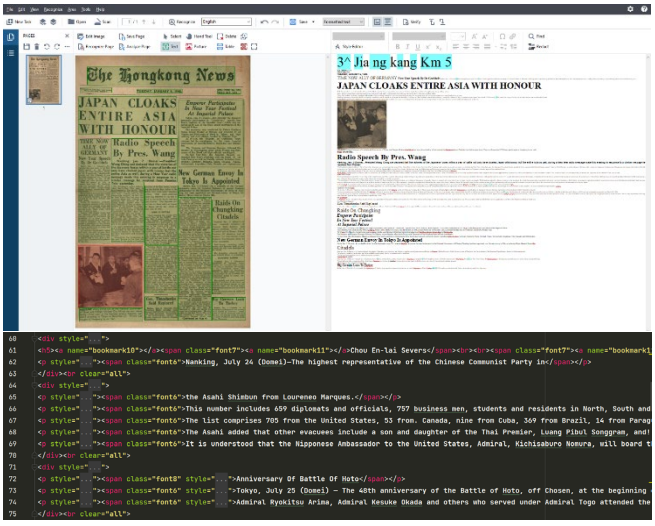

***Figure 1**. The digital image was converted to text content by ABBYY FineReader (top). OCR result has been extracted to HTML format (bottom).*

Since the extracted files were in HTML format as shown in Figure 1, our team applied conventional web scraping techniques to solve the above problem. First, we prepared around 500 sample lines as our test dataset. The dataset had around 400 noisy and 100 headline-related objects. Headline related object has been treated as a true sample and noisy has been treated as a false sample. Two main features have been considered to distinguish these true and false samples: 1) "Font Size" refers to font size in class; 2) "Length of Line" refers to the number of characters in each line.

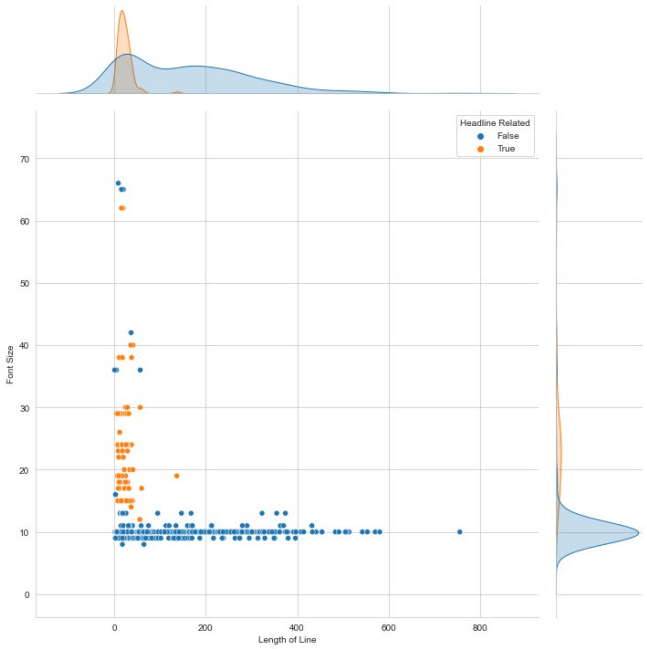

***Figure 2**. Distribution of test data by font size and the number of characters. Orange dots represented that the line content is headline-related (True sample). Blue dots represented that the line content is background noise (False sample).*

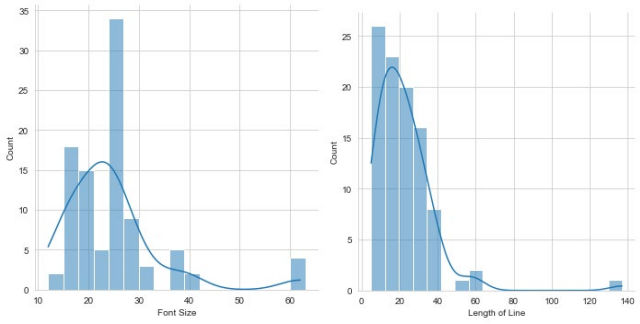

***Figure 3**. Distribution of headline-related objects (True sample) by font size and the number of characters. Compared to font size, the number of characters is a significant feature to identify a true sample.*

Our test results has been shown in figure 2 and 3. Our team found that true samples were mainly distributed between 12 – 29 characters and false samples were concentrated at around 10 font size. Based on this result, our team used Beautiful Soup, a web-scraping Python library, to extract headline from HTML files. Considering our real situation, we trade off some precision rate to increase recalled items. Based on the findings in the distribution of Font Size and Length of Line as illustrated in the above figure, our team decided to retrieve lines between 1 – 80 characters and excluded noise by filtering lines less than 12 font size. After the above steps, headline-related objects would automatically be extracted to separate files as shown in figure 4.

```
1  New Year Speech By Dr Goebbels
2  Gen. Timoshenko Said Replaced
3  Big German Loan To Turkev
4  Emperor Participates
5  In New Year Festival
6  At Imperial Palace
7  TIME NOW ALLY OF GERMANY
8  New German Envoy In Tokyo Is Appointed
9  Raids On Chungking
10 Citadels
11 Radio Speech By Pres. Wang
12 JAPAN CLOAKS ENTIRE ASIA WITH HONOUR
```

***Figure 4**. Headline related object has been extracted to the text file.*

(2) Self-developed YOLOv5 model

In strategy two, we focus on training a YOLO object detector. Due to the use of the convolutional neural network (CNN), lots of improvements have been made in object detection in recent years. YOLO is one of the famous object detectors based on this network. With the friendly deployment environment provided by YOLOv5, the custom model could be easily trained to detect a real object. Therefore, our team tried to adopt this technology to the project. By training a custom detector, we hope to automatically extract headline from images. It contained three main steps:

- Using Roboflow to annotate training data.
- Developing a YOLOv5 object detector model.
- Recognize text from cropped images through the OCR toolkit.

Our team prepared 2% of the whole data as a training set. Using Roboflow, an online software, we then annotated headline and image. To reduce overfitting, we added some newspaper images in the Roboflow universe to our training set. With reference to figure 5, the annotated headline in our training data was wide shaped and concentrated at the top space. This distribution was suitable for our real situation.

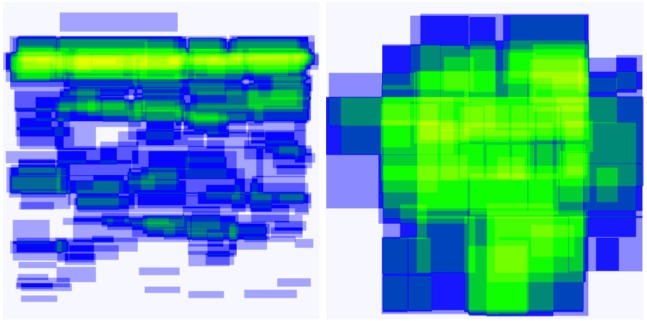

***Figure 5**. Annotation heatmap of headline (left) and image (right).*

After preparing training data, we extended the model provided by the YOLOv5 contributor in GitHub [16] and trained the object detector at Google Colab. Our training data was relatively small compared to other business object detection missions. Though the training data was inadequate, we surprisingly obtained an acceptable result. Our object detector has 0.88 mean Average Precision (mAP). Precision and recall rates were 0.97 and 0.94 respectively. Precision, recall and confidence trade-off could be found in figure 6.

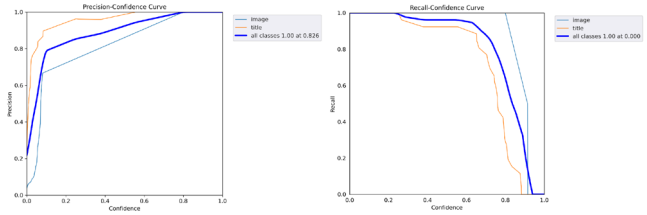

***Figure 6**. Precision, recall and confidence trade-off. It shows the precision-confidence curve (left) and recall-confidence curve(right).*

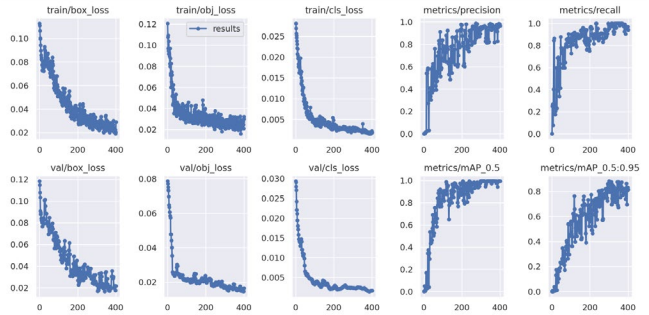

***Figure 7**. Loss and accuracy in 400 epochs. Epoch 301 has been chosen as our model since it obtained the best weight in mAP.*

The model improved swiftly in terms of precision, recall and mean average precision before plateauing after about 200 epochs. The box, objectness and classification losses of the validation data also showed a rapid decline until around epoch 250. Though we trained the model for 400 epochs, we used early stopping to select the best weight in mAP at epoch 301. Compare with precision, our team hoped to recall more objects so as to include more headlines for further analysis. Therefore, we traded off some precision rate and chose 0.2 confidence (as shown in Figure 6) to extract headline and image. Besides, we found high resolution input could significantly improve accuracy in image extraction task. Therefore, we extracted the image separately and increased the input size to 2560px. After adjusting parameters in our detection model, we cropped headline and image by calling the pre-built function in YOLOv5 and recognizing text by Tesseract.

***Figure 8**. Performance of our detection model in validation data.*

Methodology Evaluation

We have implemented the headline extraction process through ABBYY FineReader and YOLO detector above. Both of them could perform extraction and recognize headline effectively. At the end of this session, we would perform an experimental comparison of these two strategies.

From a performance aspect, both strategies have high accuracy in detecting headline and image. Since there were two different detection targets, our team would evaluate them separately. For the image target, the YOLO detector performed better than ABBYY FineReader. Their scores have been shown in figure 9. YOLO detector obtained 0.82 precision and 0.95 recall rate. On the other hand, ABBYY FineReader obtained 0.68 precision and 0.71 recall rate.

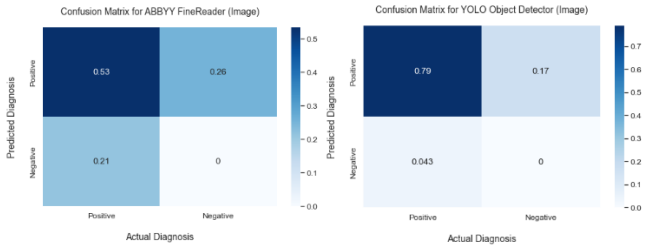

***Figure 9.** Compare the performance of ABBYY FineReader and YOLO detector in image target. Since we are evaluating an object detection task, true negative refers to the true background and it is not applicable.*

False positive (FP) was a major type of error in our image detection task. As mentioned by [17], there were different kinds of false positives which likely require different kinds of solutions. In our test dataset, false positives were mainly coming from localization (LOC) and background error (BG).

In our project, localization error would be defined as the misalignment of bounding boxes. The main reason for the LOC error of ABBYY was the software wrongly predict the space as text inside an image, as shown in figure 10. LOC error of YOLO mainly came from the specific size of the image. Our trained model had good performance in detecting small images. However, it performed comparably not well in middle-size targets. Our team guessed this error may probably cause by the imbalance of training data.

Another main false positive was the BG error. BG error means the detector was confused with the background object. ABBYY performed comparably weakly in this category. It would sometimes confuse calligraphy letters and identified them as images. It resulted in a low precision rate. On the other hand, YOLO rarely had this kind of error.

Figure 10. *ABBYY wrongly predicted the area of red rectangles as texts. (left). YOLO detector could not detect the whole image (right). ABBYY wrongly predicted the newspaper name as an image and caused a BG error. (bottom)*

False Negative (FN) was another type of error. It means our strategy wrongly detects the target as background. In our project, false negatives always occurred when our model or software identify multiple objects as one target as shown in figure 11. Although both strategies could not always distinguish close objects, YOLO could perform this in small images.

Figure 11. *Our project aimed to extract the above images as eight different images. However, ABBYY could only return a single big image (left). YOLO detector could separate them as different images (right).*

In our project, FN error is relatively important because we hope to have a high recall rate. LOC error is also important since extracted image could not be used when this error occurred. BG error is acceptable as we could delete useless images manually.

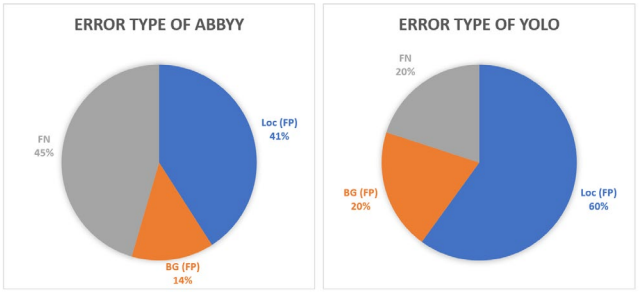

Figure 12. *Compare the error type ratio of ABBYY FineReader and YOLO detector in the image target. FN referred to our strategy as wrongly detecting the target as background. BG (FP) referred to the confusion of the background object. LOC (FP) referred to the misalignment of bounding boxes.*

For the headline target, ABBYY FineReader performed better than the YOLO detector. Their scores have been shown in figure 13. ABBYY FineReader obtained 0.74 precision and 0.98 recall rate. On the other hand, the YOLO detector obtained 0.71 precision and 0.94 recall rate. Both strategies could identify most headline and obtain a high recall rate.

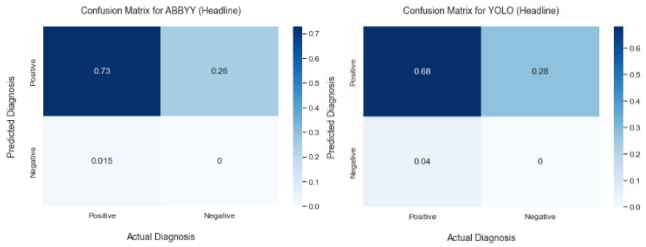

Figure 13. *Compare the performance of ABBYY FineReader and YOLO detector in headline target. Since we are evaluating an object detection task, true negative refers to the true background and it is not applicable.*

False positive is the main error in both strategies, but the error type of them was different. For the YOLO detector, LOC error was the main occurred error. Although our model performed well in the small headline, it could not accurately place the bounding box in the large one. This LOC error is probably related to the one-stage model structure of YOLO [18] and lacking large headline samples in the training dataset. Those offset of bounding boxes would cause low accuracy in the OCR process afterwards.

Figure 14. *Localization error is significant in the large headline. The headline in the upper right corner was hardly identified by OCR software since some features were cropped by a bounding box.*

For ABBYY FineReader, the BG error was the main error type. Though we used the web scraping technique to extract headline from raw HTML, some noises remained. This situation mostly happened on pages with complicated structures. Most of these noises were date and newspaper volume. Besides, the same headline of the newspaper was separated into multiple lines as shown in figure 15. It was complicated to align them automatically. Therefore, further manual post-processing is necessary and results in high labour costs.

```
1  ■ -W®
2  " NEWSTT"
3  Cl INDAY APR 11 1 1 1G12
4  Nine Enemy Planes
5  Set On Fire
6  Actual Surrender In Near Future
7  Anglo-U.S. Forces Foiled In Tunisia
```

Figure 15. *Background error of ABBYY. The headline "Nine Enemy Planes Set On fire" was separated into two lines. Further manual post-processing is necessary. Some noises could also be found in lines 1 to line 3.*

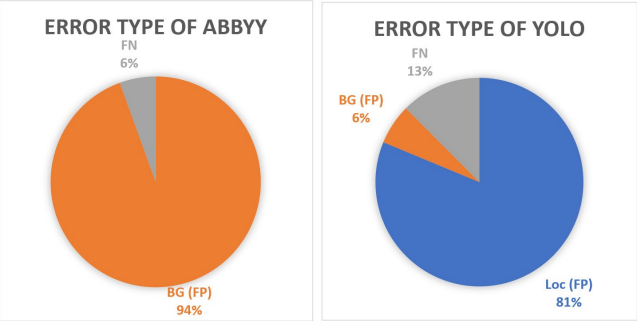

Figure 16. *Compare the error type ratio of ABBYY FineReader and YOLO detector in headline target. FN referred to our strategy as wrongly detect targets as background. BG (FP) referred to the confusion of background objects. LOC (FP) referred to the misalignment of bounding boxes.*

From time-consuming aspect, training an object detection model would spend some time, especially in data labeling process. In contrast to the YOLO detector, ABBYY FineReader does not need to spend time in the training model, but it took time in data post-processing. In our case, two weeks were spent processing 530 images.

From flexibility and deployment aspects, ABBYY FineReader is commercial software. With the friendly user interface, it provided an easy-to-use environment for text recognition. In contrast to the convenient environment of ABBYY FineReader, the YOLO detector has another advantage. Since it was open source, there were many ways and spaces for further enhancement. In our case, the bounding box problem could probably be solved by feeding more data, altering the bounding box size etc. On the other hand, ABBYY FineReader could only be improved by modifying the scraping code. After considering various perspectives of these two strategies, our team suggested that YOLO Detector could be used in larger-scale projects. For small-scale works, we recommend using ABBYY FineReader, since post-processing was acceptable in a relatively small dataset. In our project, ABBYY FineReader has been selected to extract headline and the YOLO object detector has been chosen to extract image due to their outweighed accuracy towards those 2 specific areas. Our team finally extracted 5,000 headlines and 200 images for our project.

Results

Two visualization products have been developed which focused on time and space respectively. For timeline visualization, we employed TimelineJS [19], to make the timeline for displaying *the Hongkong News* headline during the Second World War. The timeline shows the headline on all first pages of *the Hongkong News* from July 1942 to July 1945. For Geodata visualization, our team employed Folium [20] as the tool for visualization. With some modifications to our data, we utilized Folium to make an interactive map with tags of places.

Figure 17*. Geodata visualization product. Users could access the digital image in our repository by simply clicking the URL in the tags.*

Conclusions

In this project, our proposed methodologies were evaluated and two visualization products were developed to showcase the feasibility of our proposed methods to serve the storytelling purpose in the settings of time and space.

In the future, our team would discover more methods for object detection and data analysis. For object detection, we hope to use similar techniques for other tasks, such as extracting the information in the table of contents. We also plan to adopt these methods in other languages. For data analysis, further analysis on sentiment, such as opinion mining [21], can be performed. Expertise with domain knowledge can be invited to extend our findings from the studies such that some more significant insights can be derived from their perspective.

References

[1] D. Bellis, The Hongkong News, 2012. Retrieved from https://gwulo.com/the-hongkong-news.
[2] A.S. Haider, R.F. Hussein, Analysing headlines as a way of downsizing news corpora: Evidence from an Arabi-English comparable corpus of newspaper articles, Digital Scholarship in the Humanities, 35 (2019) 826-844.
[3] V.W.Y. Lum, M.K.F. Yip, Headline Analysis with Machine Learning on The Hongkong News, CUHK Research Data Repository, 2022. Retrieved from: https://doi.org/10.48668/E6JEKD.
[4] The Chinese University of Hong Kong Library, Hong Kong Tabloid Newspapers, 2022. Retrieved from https://repository.lib.cuhk.edu.hk/en/item/cuhk-2623246-0.
[5] V.W.Y. Lum, M.K.F. Yip, Headline Analysis with Machine Learning on The Hongkong News, 2023. Retrieved from https://dsprojects.lib.cuhk.edu.hk/en/projects/heading-analysis-machine-learning-hongkong-news/tabloid-hknews-geodata-visualization/.
[6] V.W.Y. Lum, K.L. YEUNG, Text Analysis and Visualisation of The Observatory Review from Hong Kong Early Tabloid Newspaper, 2022. Retrieved from https://dsprojects.lib.cuhk.edu.hk/en/projects/hong-kong-early-tabloid-newspapers/tabloid-introduction/.
[7] E.A. Msuya, Analysis of Newspaper Headlines: A Case of Two Tanzanian English Dailies, Journal of Education, Humanities, and Sciences, 8 (2019).
[8] C. Develotte, E. Rechniewski, Discourse analysis of newspaper headlines : a methodological framework for research into national representations, (2001).
[9] T. Fogec, Critical Discourse Analysis of Tabloid Headlines, 2014.
[10] N. Aqromi, An Analysis of Metaphor for Corona on Headlines News, PIONEER: Journal of Language and Literature, 12 (2020) 157.
[11] M. Arshad, N. Khan, A critical discourse analysis of the Pakistani newspaper headlines on the federal budget for FY 2021-2022, Journal of Humanities, Social and Management Sciences (JHSMS), 2 (2021) 176-186.
[12] D. Dor, On Newspaper Headlines as Relevance Optimizers, Journal of Pragmatics, 35 (2003) 695-721.
[13] R. Saha, A. Mondal, C.V. Jawahar, Graphical Object Detection in Document Images, The Institute of Electrical and Electronics Engineers, Inc. (IEEE), Piscataway, 2019, pp. 51-58.
[14] X. Yi, L. Gao, Y. Liao, X. Zhang, R. Liu, Z. Jiang, CNN Based Page Object Detection in Document Images, The Institute of Electrical and Electronics Engineers, Inc. (IEEE), Piscataway, 2017, pp. 230-235.
[15] B.C.G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, D.S. Weld, The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America, pp. 3055-3062.
[16] Ultralytics, YOLOv5 Github documentation. Retrieved from https://github.com/ultralytics/yolov5.
[17] D. Hoiem, Y. Chodpathumwan, Q. Dai, Diagnosing Error in Object Detectors, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 340-353.
[18] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, IEEE, pp. 779-788.
[19] Knight lab, TimelineJS, 2023. Retrieved from https://timeline.knightlab.com/.
[20] Rob Story, Folium documentation, 2013. Retrieved from https://python-visualization.github.io/folium/.
[21] J.R. Chaudhary, J. Paulose, Opinion mining on newspaper headlines using SVM and NLP, International Journal of Electrical and Computer Engineering, 9 (2019) 2152-2163.

Author Biography

Michael YIP received his M.Phil. in History from The Chinese University of Hong Kong. He has previously worked as Research Assistant to support digital scholarship exercises for the Chinese University of Hong Kong Library. He is now working as Project Coordinator at the same unit focusing on digitalization, project management and exploring the new ways to expose the data in Digital Repository to support various digital scholarship activities

Vincent LUM received his M.Phil. in Computer Science and Information Systems from The University of Hong Kong. His research interest was in the area of Ubiquitous and Pervasive computing. He has joined the medical library in Hospital Authority of Hong Kong to execute infrastructural development in Library Services Platform. He is now working at the Chinese University of Hong Kong Library and is responsible for developing the Library's digital initiatives to maximize the effective use of emerging digital technologies. His duty also includes digital services and digitization to advocate the use of Digital Repository to incorporate digital scholarship tools into these digital collections to support digital scholarship research.