ARTICLE

Unlocking the Digitized Historical Newspaper Archive

Exploring Historical Insights with Deep Learning

Vincent Wai-Yip Lum and Michael Kin-Fu Yip

ABSTRACT

This paper aims to utilize historical newspapers through the application of computer vision and machine/deep learning to extract the headlines and illustrations from newspapers for storytelling. This endeavor seeks to unlock the historical knowledge embedded within newspaper contents while simultaneously utilizing cutting-edge methodological paradigms for research in the digital humanities (DH) realm. We targeted to provide another facet apart from the traditional search or browse interfaces and incorporated those DH tools with place- and time-based visualizations. Experimental results showed our proposed methodologies in OCR (optical character recognition) with scraping and deep learning object detection models can be used to extract the necessary textual and image content for more sophisticated analysis. Timeline and geodata visualization products were developed to facilitate a comprehensive exploration of our historical newspaper data. The timeline-based tool spanned the period from July 1942 to July 1945, enabling users to explore the evolving narratives through the lens of daily headlines. The interactive geographical tool can enable users to identify geographic hotspots and patterns. Combining both products can enrich users' understanding of the events and narratives unfolding across time and space.

INTRODUCTION

Historical newspapers are now preserved and made accessible through digital repositories, which have transformed the way we work with and appreciate these valuable resources, making our heritage and historical records more accessible than ever before. Their availability empowers researchers to unlock profound knowledge, providing an important insight to the multifaceted society and cultural circulation at the time and broadening the horizons of digital humanities research and its practical applications. The generation of extensive digitized newspaper collections has furnished researchers across diverse domains with a powerful tool to advance their investigative endeavors.

Contemporary digital humanities tools empower scholars with multifaceted avenues to explore the intricacies of the past. They can navigate through interactive timelines, delve into vast repositories of digitized textual content, or unravel the narratives of significant events through newspaper headlines.

This project embarks on an exploration of the interplay between narrative time and space. It embraces a conception of geography and space as loci where contemporary lives intertwined with the narratives encapsulated within newspaper articles. Time, on the other hand, emerges as a fundamental axis of storytelling, enabling digital scholarship to unfold through the lens of temporality. By harnessing sophisticated text mining computational algorithms, this research

About the Authors

Vincent Wai-Yip Lum (vincentlum@cuhk.edu.hk) (corresponding author) is Digital Technologies Librarian, The Chinese University of Hong Kong. **Michael Kin-Fu Yip** (michaelyip@hsu.edu.hk) is Collection and Programme Support Librarian, The Hang Seng University of Hong Kong Library.

Submitted: 25 November 2024. Accepted for Publication: 16 July 2025. Published: 15 September 2025.

https://doi.org/10.5860/ital.v44i3.17292

endeavors to unveil novel ways in which digital tools can facilitate an understanding of newspaper content through the prisms of time and space, allowing geography to intersect with history.

The overarching objective is to empower scholars with multidimensional access to content, fostering a nuanced comprehension of newspaper articles through unconventional approaches. The Hong Kong Early Tabloid Newspapers《香港早期小報》collection, launched in 2022, serves as a testament to this endeavor. Encompassing tabloid newspapers published in Hong Kong during the twentieth century, this collection offers a lens into the leisure and entertainment of the masses, covering a diverse array of topics ranging from politics and operas to dramas, comics, and even pornography.

Within the context of this project, *The Hongkong News*, a pivotal holding from the Hong Kong Early Tabloid Newspapers Collection, is subjected to the aforementioned approaches. The Japanese Occupation edition of *The Hongkong News*, which commenced publication on December 25, 1941, immediately following the British Crown Colony's surrender, and continued until August 17, 1945, the week preceding Hong Kong's liberation. It provides scholars with a unique perspective—the voice of Japan from within Hong Kong. The historical significance of this newspaper has been underscored in previous studies, highlighting its invaluable contribution to scholarly discourse.[1]

Through the extraction of headlines from *The Hongkong News*, two visualization products were developed, offering tangible outcomes of this research endeavor. Employing this project as a case study, we share the experiences and insights garnered while navigating through these newspaper headlines. The proposed holistic framework encompasses the following pivotal components:

- Extracting visual and textual content from historical newspapers.
- Applying deep learning algorithms for headline extraction.
- Extracting data through computational techniques, such as natural language processing (NLP) and geocoding.
- Developing two visualization products that facilitate public and scholarly access to our existing repository.

These techniques prove instrumental in visualizing the historical knowledge and entities embedded within the newspaper contents. Headline analysis serves as a qualitative analysis of the full news stories and has proven to be an effective "down-sampling" approach. This approach refers to downsizing a large corpus of newspaper articles and analyzing them using the headline as the qualitative topic with a higher level of abstraction. This can reduce vast news corpora to a manageable dataset.

MOTIVATION

The Hongkong News serves as a rich repository of information, captivating the interest of scholars. However, their exploration is currently constrained by the need for manual page-by-page searches. For instance, when users seek to uncover headlines mentioning specific locations, they are compelled to browse through thousands of pages within our repository. Furthermore, the existing digital assets possessed untapped potential to provide value beyond mere preservation and online access, underscoring the opportunities for optimizing the utilization of these digitized images. Transcending the confines of the image itself, our team attempted to apply various computational techniques to unlock the multifaceted value embedded within this tabloid newspaper. Another challenge confronting users was the ability to derive concrete insights from the newspaper content, particularly when the corpus in question was of a substantial scale. We

aimed to address this challenge by performing headline analysis, a tool recognized for its efficacy in newspaper topic analytics, as echoed by Haider and Hussein.[2]

To further promote the multifaceted content of tabloid newspapers and enhance their accessibility, our project pursued two primary objectives. First, we provided a semiautomatic procedure to recognize and extract newspaper headlines. Second, we developed two visualization products to facilitate user access to our digital repository. These visualization products not only afforded scholars insights into headlines appearing across different times and spaces but also enabled them to address humanities-related inquiries, such as "Which battle was the focus of Japanese propaganda at a specific time?" and "How did Japanese propaganda evolve during the war?"

The two visualization products can showcase the temporal (through the timeline visualization) and spatial dimensions (through the geodata visualization) of the information. For timeline visualization, we created a timeline-based tool exploring the headlines of the news, with illustrative images extracted from the newspaper pages using an object detection computational algorithm. These images can serve as a tangible glimpse into the historical events spanning the period of the Second World War. For geodata visualization, an interactive geographical tool was created with the extracted place names from the newspaper pages. To showcase the distribution of places mentioned throughout the newspaper headlines, heatmap visualizations can be featured in addition to the spatial exploration. This can be beneficial to the users to explore the geographic hotspots and patterns, further boosting their understanding of the newspaper content through multifaceted visualizations.

This project encompassed 530 issues of *The Hongkong News* spanning the years 1942 to 1945. The related dataset is available at the CUHK Research Data Repository, while the digital images of *The Hongkong News* can be accessed through our Digital Repository. The visualization product, with sample source code also available on GitHub, can be found on our Digital Scholarship Project.[3]

RELATED WORK

Headline Analysis

The advent of digitization and optical character recognition (OCR) has paved the way for in-depth textual analysis. Most studies in newspaper analysis focus on textual analysis at the article level. Typically, headline analysis in newspapers is one such tool. For instance, numerous researchers have employed headline analysis as a qualitative analysis for full news stories in newspapers.[4] The headline, being a crucial component of a news story and capturing the reader's attention through its summary, has been found to produce complementary and convergent findings with corpus analysis. In one study, the author concluded that analyzing headlines proved to be an effective "down-sampling" option for reducing large news corpora to a manageable dataset.[2] However, most studies still employed conventional methods, and when the newspaper corpus was of a substantial scale, this posed a challenge. Automatic or semi-automatic means of headline extraction, therefore, became necessary.

Object Extraction

Conventionally, given the varying layouts and styles of printed materials, paper headlines could only be extracted manually. Without computational support, the workflow was time-consuming and labor-intensive. Although some commercial OCR software could aid in text recognition, the effectiveness of OCR varies considerably, especially when applied to historical newspapers, where the accuracy can be influenced by factors such as print quality, paper condition, and layout

complexity. For instance, studies highlight that while OCR can achieve over 98% accuracy in optimal conditions, historical newspapers often see much lower rates.[5]

Additionally, despite the support provided by OCR software for text recognition, researchers still had to manually search for and retrieve headlines from the OCR results, as headlines could not be extracted automatically. With the advent of convolutional neural networks (CNNs), significant improvements have been made in object detection in recent years. Numerous object detectors have been developed to address user demands. Object detection has been employed to automatically detect different illustrations in modern publications.[6]

In 2020, the Library of Congress applied the Faster-RCNN model to extract information from newspaper pages in their broad-scale project.[7] Their training dataset contained up to 3,600 pages with 48,409 annotations. The model was trained for 17 hours on the NVIDIA T4 GPU. Although the precision of headline extraction was near 75% in average precision in the validation set, this model could not generally fit other data, exhibiting a relatively low generalization ability. Given that the visual content recognition model was trained on World War I–era newspapers, repurposing the model to nineteenth-century newspapers resulted in a performance dropoff in average precision (AP): (AP: 21.2% for 1850–1875 newspapers and AP: 51.6% for 1875–1900 newspapers). It proved challenging to repurpose the model to other corpora due to the nongeneric nature of the dataset. Additionally, while their studies employed the Detectron2 model, we utilized the YOLO recognition model to explore potential performance improvements. YOLO (You Only Look Once) is a prominent object detection algorithm that leverages the deep learning techniques.

Through this comprehensive exploration, encompassing headline analysis, object extraction techniques, and the development of visualization products, our project aimed to unlock the rich insights embedded within *The Hongkong News*, fostering enhanced accessibility and scholarly engagement with this invaluable historical resource.

HEADLINES AND IMAGES: A TWO-PRONGED APPROACH

To facilitate the extraction of headlines and images from the historical newspaper corpus, two distinct methodologies were explored and evaluated: (1) leveraging optical character recognition (OCR) software (ABBYY FineReader) in conjunction with web scraping techniques, and (2) developing a custom object detection model based on the YOLOv5 architecture. By contrasting these two approaches, we aimed to establish a robust framework for headline and image extraction, tailored to the unique characteristics of the data at hand.

OCR-Driven Extraction and Web Scraping

In the first methodology, we harnessed the power of OCR technology, specifically the widely acclaimed ABBYY FineReader software, to convert the digital images of the newspaper into machine-readable HTML formats. Although ABBYY FineReader excels at text recognition, it lacks dedicated functionality for extracting specific content elements, such as headlines. To address this limitation, we employed web scraping techniques using the Python library Beautiful Soup. This two-step process involved

1. Utilizing ABBYY FineReader to extract the textual content from the digital images and represent it in HTML format.
2. Applying web scraping techniques through Beautiful Soup to parse the HTML and isolate the headline content.

The HTML output from ABBYY FineReader presented a commingled representation of all textual elements within the digital image (as shown in fig. 1), necessitating further data processing to differentiate headlines from other content types. To facilitate this distinction, we curated a test dataset comprising approximately 500 sample lines, with 400 representing noise and 100 containing headline-related content. Two key features were leveraged to distinguish between true (headline) and false (noise) samples: font size in and line length.

Figure 1. Utilizing ABBYY FineReader to extract the textual content from the digital images (top) and represent it in HTML format (bottom).

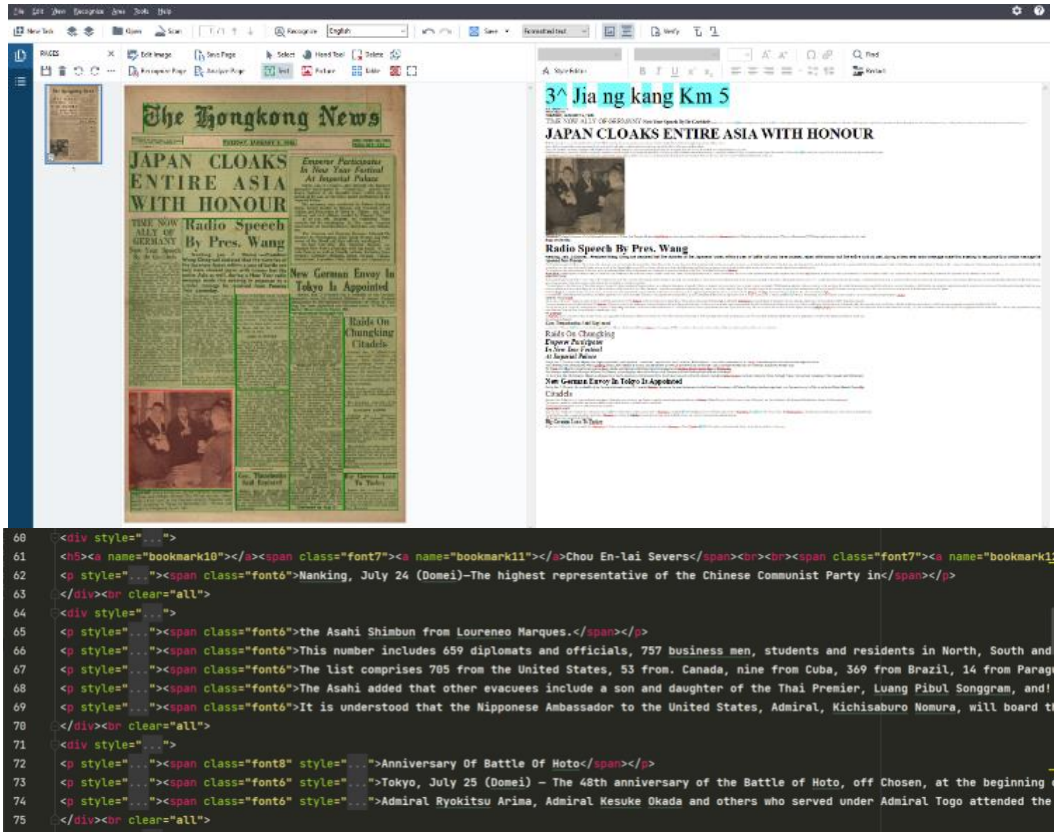


Figure 2. Distribution of test data by font size and the number of characters. Orange dots represented that the line content is headline-related (true sample). Blue dots represented that the line content is background noise (false sample).

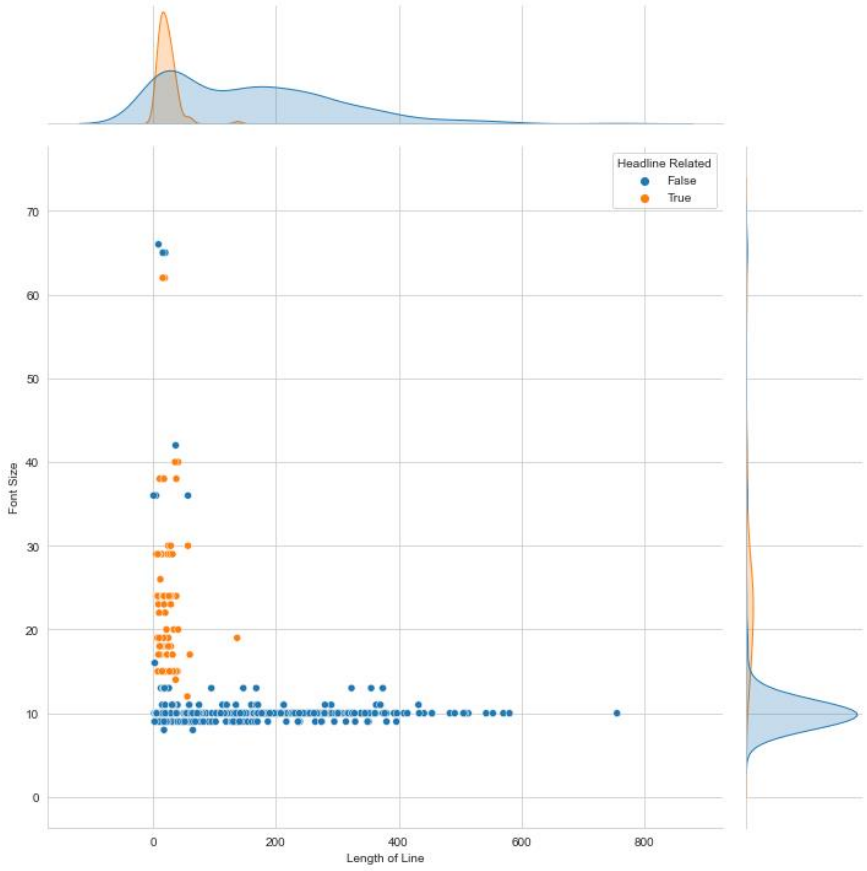
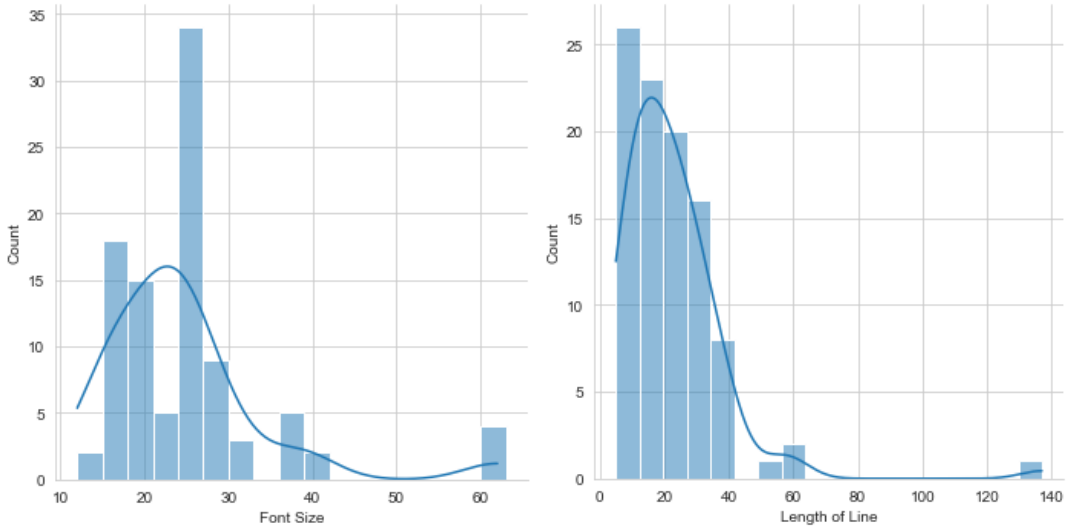


Figure 3. Distribution of headline-related objects (true sample) by font size and the number of characters. Compared to font size, the number of characters is a significant feature to identify a true sample.



Our analysis of the test data revealed that true samples predominantly fell within a character range of 12 to 29 as shown in figures 2 and 3, while false samples tended to cluster around a font size of 10. Guided by these insights, we employed Beautiful Soup to extract lines ranging from 1 to 80 characters in length and filtered out lines with a font size below 12, effectively isolating the headline-related content (see fig. 4).

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

```
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
```

Customed YOLOv5 model

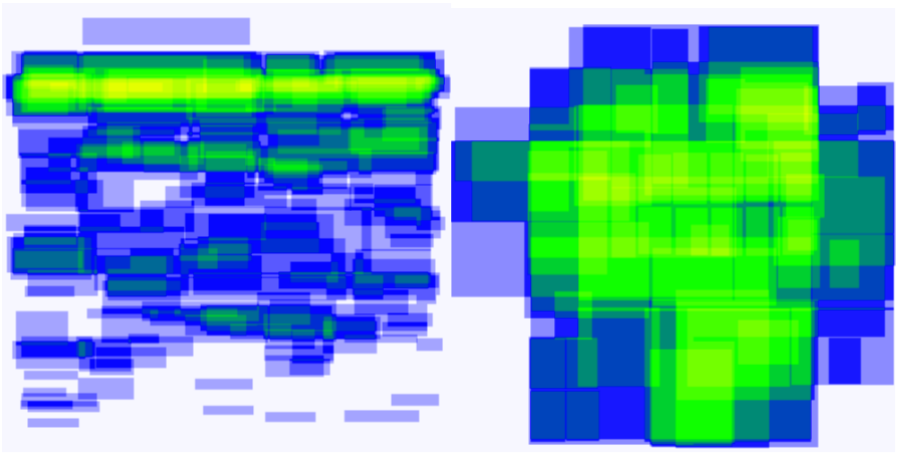
The second methodology involved developing a tailored object detection model using the YOLOv5 architecture. Recent years have witnessed significant advancements in object detection, driven by the adoption of convolutional neural networks (CNNs). The YOLOv5 framework provides a user-friendly environment for training and deploying custom object detectors, making it well-suited for our project's requirements.

Our team pursued this approach with the goal of automatically extracting headlines and images from the newspaper corpus. The development process encompassed three primary steps:

1. Annotating training data using Roboflow.
2. Training a YOLOv5 object detector model.
3. Recognizing text from cropped images using an OCR toolkit.

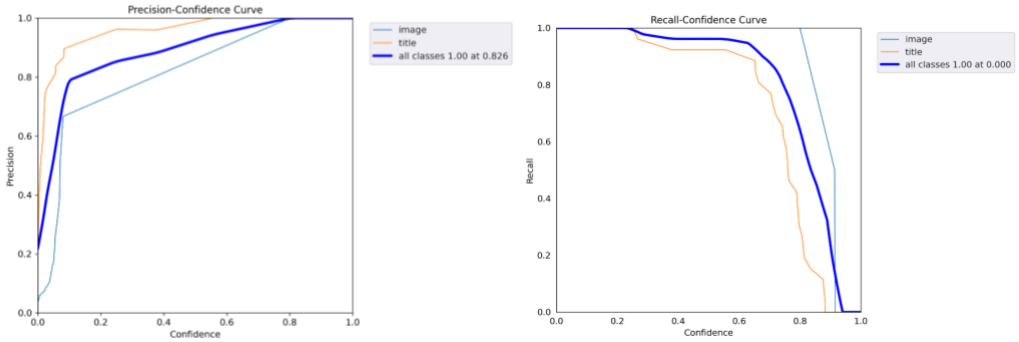
To construct the training dataset, we utilized 2% of the entire data corpus. Leveraging Roboflow, an online annotation tool, we meticulously labelled headlines and images within the newspaper pages. To enhance the model's generalization capabilities and mitigate overfitting, we augmented the training set with additional newspaper images from the Roboflow universe. The annotated headlines exhibited a distinctively wide shape concentrated in the upper regions of the page, aligning with the typical layout observed in our data (see fig. 5).

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



With the training data prepared, we extended the YOLOv5 model architecture provided by the project's contributors on GitHub and trained the object detector on Google Colab. Despite the relatively modest size of our training dataset compared to typical object detection tasks, we achieved remarkably accurate results. Our custom detector attained a mean average precision (mAP) of 0.88, with precision and recall rates of 0.97 and 0.94, respectively, as depicted in figure 6.

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



The performance of our custom YOLOv5 object detector on the validation data is illustrated in figure 7, showcasing its efficacy in accurately detecting and localizing headlines and images within the historical newspaper corpus. Through this approach, we successfully developed a tailored object detection model capable of automatically extracting headlines and images, paving the way for further analysis and visualization of these invaluable historical resources.

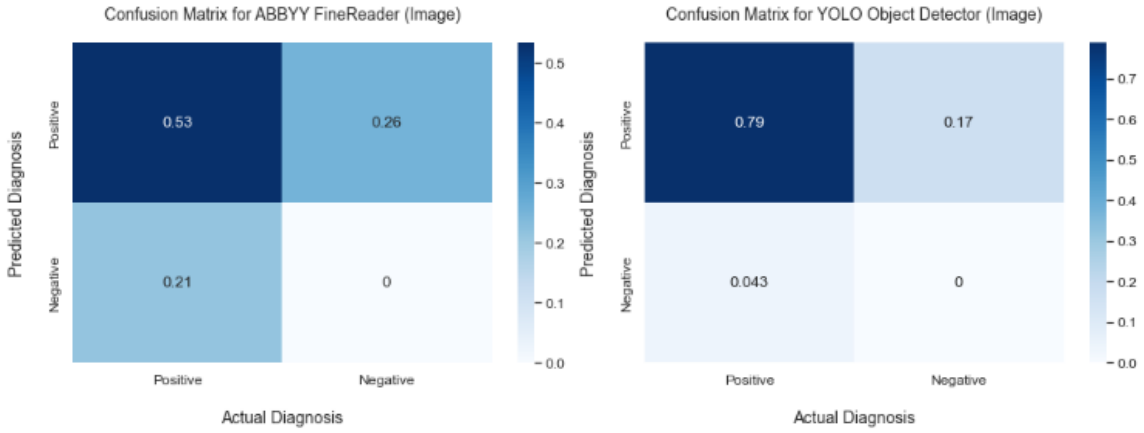
Figure 7. Performance of our detection model in validation data.

METHODOLOGY EVALUATION

In pursuit of extracting headlines and images from our historical newspaper archive, we implemented and evaluated two distinct methodological frameworks: optical character recognition (OCR) coupled with web scraping techniques, and a custom object detection model underpinned by the YOLOv5 deep learning architecture. A comprehensive experimental comparison was undertaken to gauge the relative performance of these approaches. Both approaches could perform extraction and recognize headlines effectively.

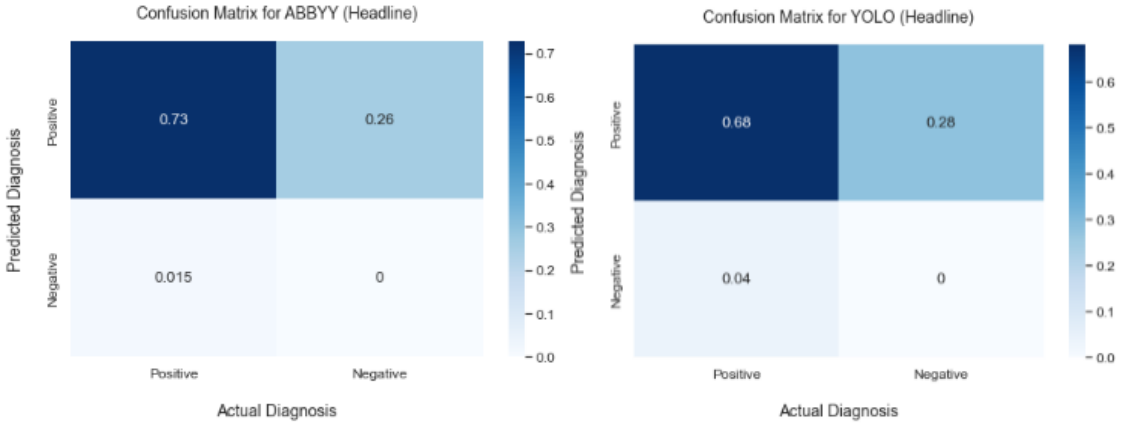
For image extraction, our findings unequivocally demonstrated the superiority of the YOLOv5 model over the ABBYY FineReader solution. As depicted in figure 8, the deep learning model attained an impressive precision of 0.82 and a recall rate of 0.95, outperforming the ABBYY FineReader's scores of 0.68 and 0.71 for precision and recall, respectively.

Figure 8. Comparing 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.



Conversely, for the task of headline extraction, the ABBYY FineReader demonstrated a distinct advantage over the deep learning model. As illustrated in figure 9, the ABBYY FineReader achieved a precision of 0.74 and an impressive recall rate of 0.98, outperforming the YOLOv5 detector, which attained scores of 0.71 and 0.94 for precision and recall, respectively.

Figure 9. Comparing 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.



From a time and resource perspective, the deep learning approach necessitated a substantial upfront investment in annotating training data and model optimization. In contrast, ABBYY FineReader did not require a dedicated training phase; however, it demanded significant post-processing efforts, as exemplified by the two-week duration required to process 530 images in our project.

Regarding flexibility and deployment, ABBYY FineReader is a commercial product with a user-friendly interface tailored for text recognition tasks. Conversely, the open-source nature of the YOLOv5 architecture offers opportunities for further customization and enhancement, such as adjusting bounding box dimensions or incorporating additional training data to address specific limitations.

After careful consideration of the relative merits and drawbacks of each strategy, our team concluded that the YOLOv5 detector is better suited for large-scale projects where the upfront investment in model training and optimization is justified. For smaller endeavors with limited data, the ABBYY FineReader may be preferable due to its ease of use and acceptable post-processing overhead.

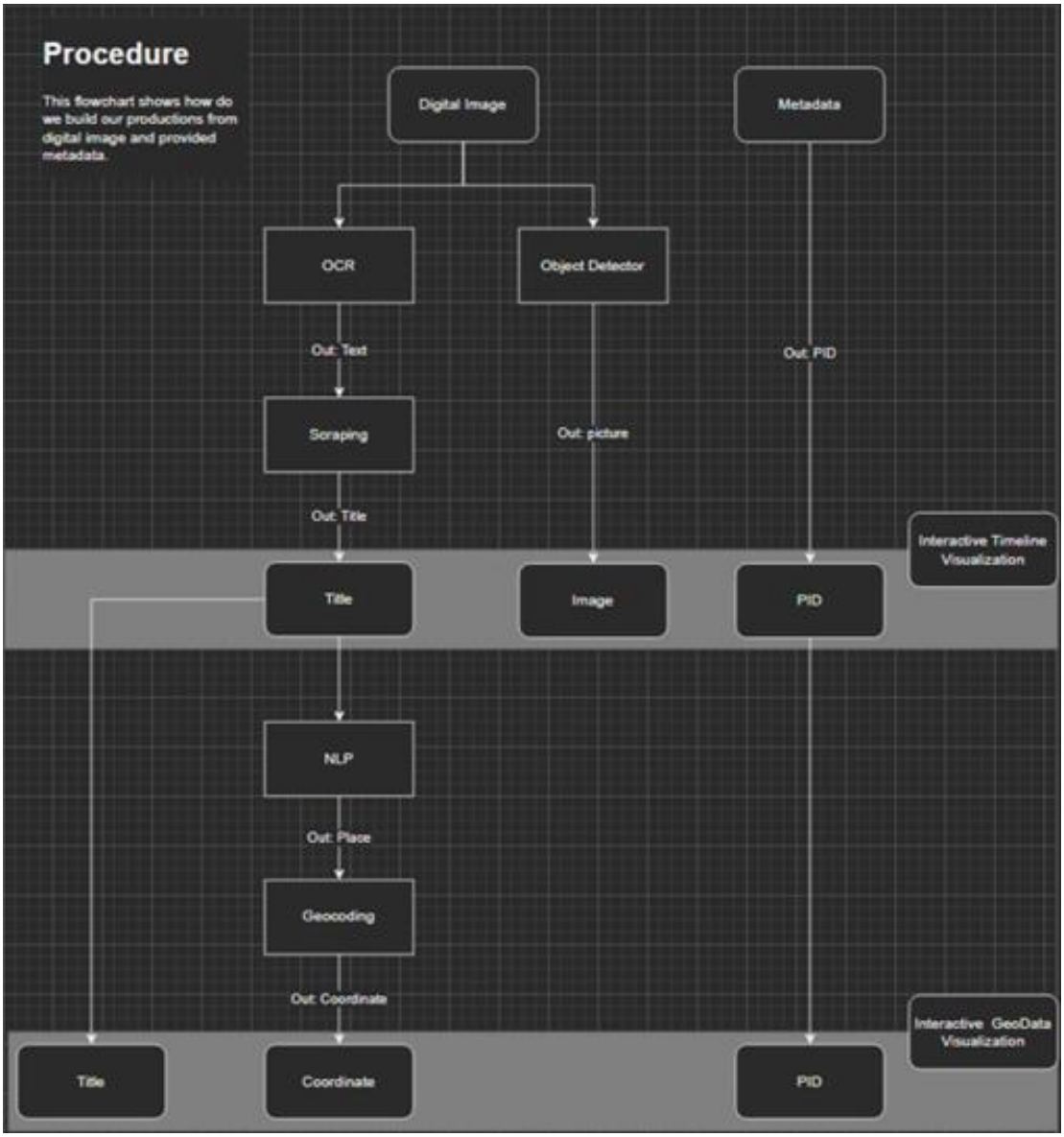
Leveraging the respective strengths of each methodology, we ultimately employed ABBYY FineReader for headline extraction and the YOLOv5 object detector for image extraction, capitalizing on their demonstrated proficiencies in these specific tasks. This integrated approach yielded 5,000 extracted headlines and 200 extracted images from our historical newspaper corpus for further analysis and visualization.

FURTHER DATA PROCESSING FOR VISUALIZATIONS

After extracting headlines from digital images, our team further processed the data by using some other text mining techniques. Our team aimed to build a geospatial visualization product. We need to identify place objects in 5,000 headlines and locate their coordinates. Traditionally, this process

could only be handled manually. It was time- and labor-consuming. However, by using some computational techniques, we could process this data automatically. Natural language processing (NLP) is the branch of artificial intelligence (AI) concerned with giving computers the ability to understand the text. Under the umbrella of NLP, Name entity recognition (NER) has attracted increasing attention. NER is a task of information extraction that seeks to locate named entities, such as persons, places, and countries. In our project, spaCy, an open-source NLP Python library, has been chosen for extracting name entities. Using NER, our team extracted up to 2,700 place objects from newspaper headlines.

Figure 10. Our project workflow from image to product.



After extracting the place objects, our team further located the coordinates by geocoding. Geocoding is the process of converting addresses into geographic presentations, which we can use to place markers on the map. In this project, Google Geocoding API was chosen to find the coordinates of the places that appeared in the newspaper. By inputting the name of a place, Google

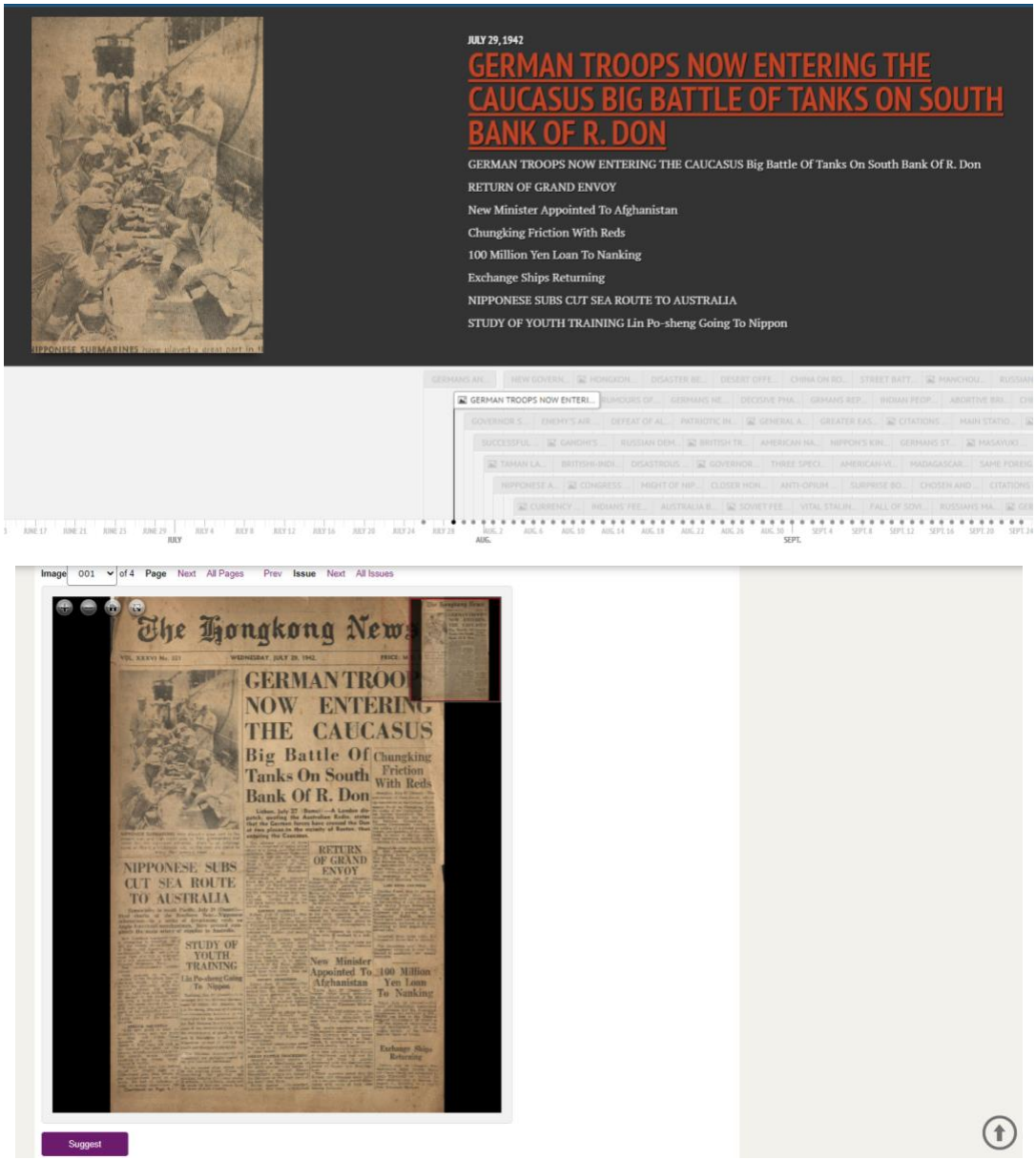
Geocoding API can return data located in the Google database. As a result, up to 400 coordinates were found.

After collecting the extracted headline, place object, and coordinates, our team paired up this information with digital images in our repository. Since many places have been found in our project, our team sorted the place list by their frequency of occurrence and chose the first 150 places for the showcase. Detailed workflow can be found in figure 10.

RESULTS: TIMELINE AND GEODATA VISUALIZATIONS

To facilitate a comprehensive exploration of our historical newspaper data, we developed two distinct visualization products, each tailored to showcase the temporal and spatial dimensions of the information.

Figure 11. Timeline visualization product (top), which is linked to the image in the repository (bottom).



For the temporal visualization, we employed TimelineJS, a powerful tool developed by Northwestern University's Knight Lab, to create an interactive timeline depicting the headlines from the front pages of *The Hongkong News* during the Second World War. This immersive timeline spanned the period from July 1942 to July 1945, enabling users to explore the evolving narratives through the lens of daily headlines. Through seamless integration with the CUHK Digital Repository, users can access the complete digitized issues of *The Hongkong News* from the Hong Kong Early Tabloid Newspapers collection by simply clicking on a specific day's headline as shown in figure 11.

To enhance the visual experience, we incorporated illustrative photographs extracted from the newspaper pages using our YOLO object detection methodology. These images provided users with a tangible glimpse into the historical events unfolding on a particular date. While not all issues contained images, we successfully extracted a total of 203 images from the newspaper corpus for inclusion in the timeline presentation.

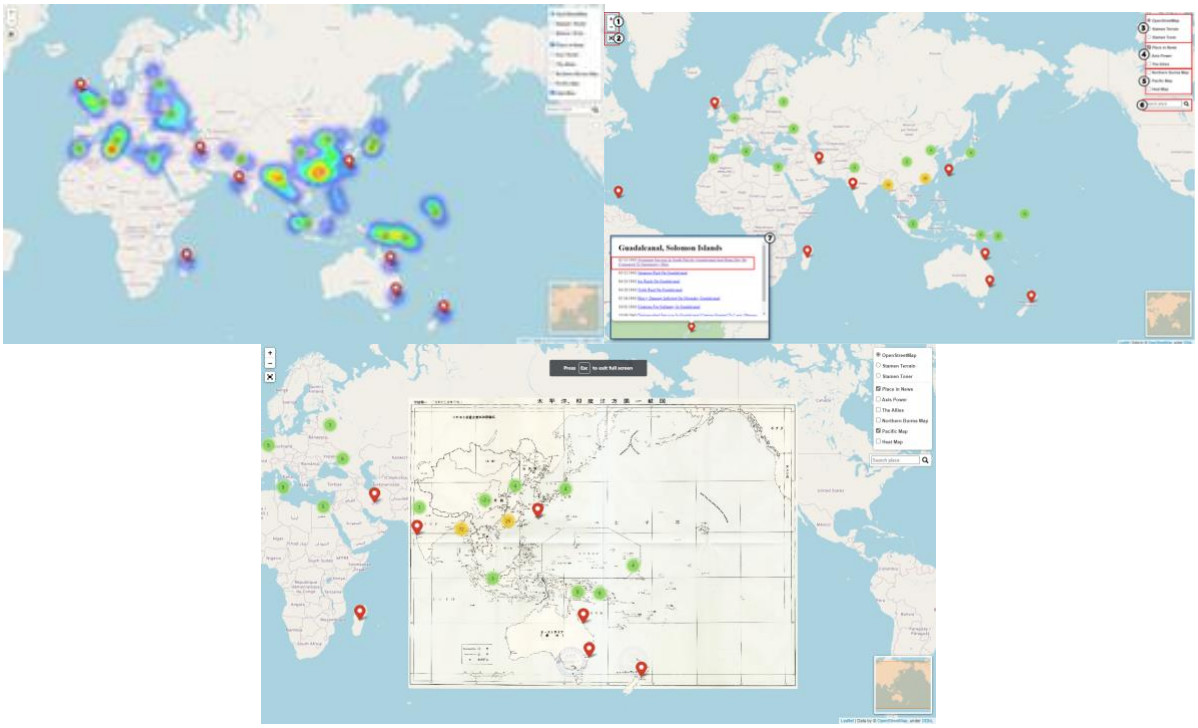
For the spatial visualization component, our team leveraged Folium, an open-source library designed for displaying geospatial data. Folium's capabilities allowed us to mark and annotate geo-political entities (GPEs), such as countries and cities, on an interactive map. Through careful data processing and integration, we utilized Folium to create an interactive map featuring tagged locations. By clicking on these tags, users can seamlessly access the corresponding digitized issues from the Hong Kong Early Tabloid Newspapers collection in the CUHK Digital Repository.

To enhance the user experience and facilitate targeted exploration, our visualization product categorized the extracted place names into three distinct categories: Places in News, Axis Powers, and The Allies. Country names, such as Japan and China, were classified as Axis Powers and The Allies, respectively, reflecting their political allegiances during the war. Headlines mentioning these countries often pertained to political matters. Place names like Yangon and Hong Kong, on the other hand, were grouped under Places in News, as these headlines typically focused on battle strategies and military operations. Users can leverage these category filters to refine their exploration and locate information of particular interest. Furthermore, we incorporated two historical maps into our visualization product, providing users with valuable context and geographical references from the newspaper's era. Recognizing the Japanese perspective of *The Hongkong News* during the Second World War, we selected maps of the Pacific and North Burma regions, which encompassed significant battlefields for Japan's military campaigns. These maps not only displayed the historical place names but also offered insights into the geographical landscape and strategic locations of that period.

To complement the spatial exploration, we integrated heatmap visualizations, including a heatmap with a temporal component, to illustrate the frequency and distribution of place mentions throughout the newspaper headlines. This feature enabled users to identify geographic hotspots and patterns, enriching their understanding of the events and narratives unfolding across time and space (see fig. 12).

Through this innovative integration of temporal and spatial visualization techniques, our product offers an immersive and multifaceted exploration of our historical newspaper data, allowing users to navigate the intricate tapestry of events, locations, and narratives that shaped this pivotal era.

Figure 12. Geodata visualization product. Users could understand the place distribution by heatmap (top left). Users could access the digital image in our repository by simply clicking the URL in the tags (top right). Users could show the historical map of our product (bottom).



FINDINGS

From GIS heatmap analysis, patterns can be identified, showcasing insights from the map and assisting in exposing possible research topics or questions. For example, our team observed that the battlefields were concentrated in some regions as shown in figure 13.

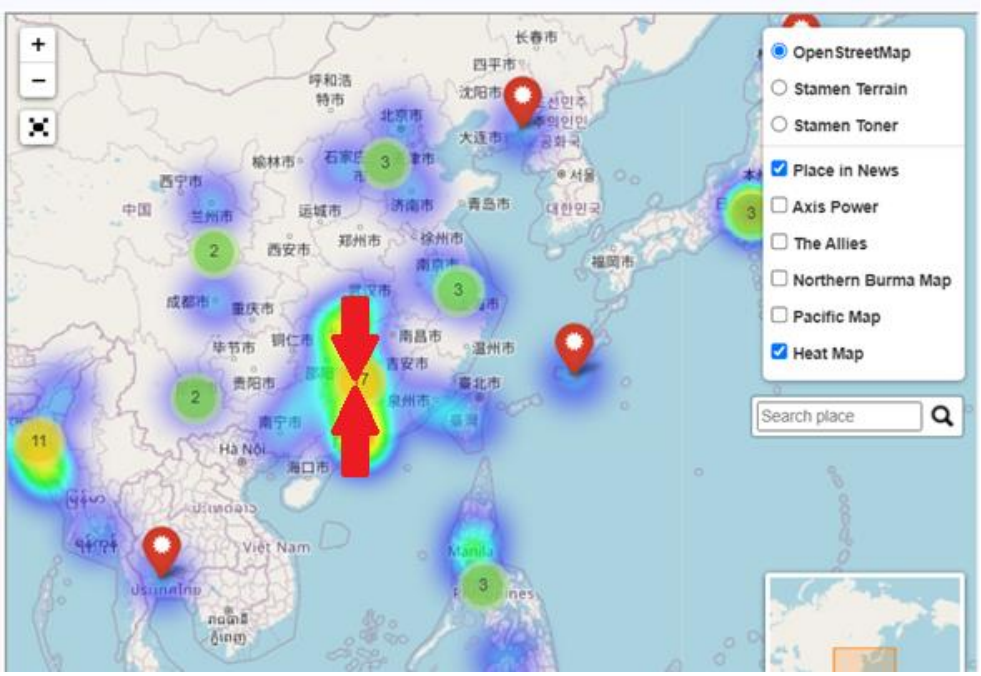
In South China, the markers were concentrated in Hunan Province. Changde (常德) is one of the important places with nine identified headlines. Most of them were related to the Battle of Changde in December 1943. Also, an implicit straight line connected from Hunan to Guangdong and Guangxi has been found. This may relate to the Tairiku Datsū Sakusen (大陸打通作戦) mindset of Japan and Operation Ichi-Go campaign (一号作戦) in 1944.

In Myanmar (Burma), the markers were concentrated in two areas, northeast India and southern Myanmar. In northeast India, 21 headlines have been found at Imphal. Most of the headlines were between March and April of 1944, which related to the Battle of Imphal (ウ号作戦). Another main area is located in southern Myanmar, including Yangon, Maungdaw, Buthidaung, and Akyab. The headlines were mainly related to the Burma campaign between 1942 and 1943.

Many markers were concentrated in the South Pacific, close to our prediction, especially in the Solomon Islands and New Guinea. Places such as Rabaul, Bougainville, Arawe, and Guadalcanal were mentioned in many headlines. Those were important battlefields in the Pacific War.

To understand more content about the distribution of the heatmap provided, researchers could further click the place name and access our repository. By accessing the newspaper images, the researcher could discover further historical content.

Figure 13. An implicit straight line connecting Hunan to Guangdong and Guangxi has been found. This may relate to the Tairiku Datsū Sakusen (大陸打通作戦) mindset of Japan.



CONCLUSIONS

Through this project, we have successfully demonstrated the feasibility and effectiveness of our proposed methodologies in enabling compelling storytelling experiences that traverse the dimensions of time and space. The two visualization products developed serve as tangible manifestations of these capabilities, showcasing the potential for extracting and presenting historical narratives in immersive and insightful ways.

Moving forward, our team is committed to exploring and integrating advanced techniques for object detection and data analysis, with the goal of continually enhancing our capabilities. In the realm of object detection, we aim to extend our methodologies to encompass tasks such as extracting information from tables of contents, while also investigating the applicability of these approaches to other languages.

On the data analysis front, we envision incorporating sentiment analysis techniques, including opinion mining, to derive deeper insights and uncover nuanced perspectives within our textual data.[8] Furthermore, we intend to collaborate with domain experts, leveraging their specialized knowledge to broaden the scope of our findings and unearth previously unexplored narratives and interpretations.

Through this iterative process of methodological refinement, interdisciplinary collaboration, and continuous innovation, we endeavor to push the boundaries of historical data visualization and storytelling, unlocking new avenues for scholarly inquiry and public engagement with our rich cultural heritage.

ENDNOTES

[1] D. Bellis, "The Hongkong News," *Gwulo: Old Hong Kong*, February 15, 2012, https://gwulo.com/the-hongkong-news.

[2] A. S. Haider and R. F. Hussein, "Analysing Headlines as a Way of Downsizing News Corpora: Evidence from an Arabic-English Comparable Corpus of Newspaper Articles," *Digital Scholarship in the Humanities* 35, no. 4 (2019): 826–44, https://doi.org/10.1093/llc/fqz074.

[3] M. K. F. Yip and V. W. Y. Lum, "Headline Analysis with Machine Learning on *The Hongkong News*," 2023, https://dsprojects.lib.cuhk.edu.hk/en/projects/heading-analysis-machine-learning-hongkong-news/tabloid-hknews-geodata-visualization/.

[4] E. A. Msuya, "Analysis of Newspaper Headlines: A Case of Two Tanzanian English Dailies," *Journal of Education, Humanities, and Sciences*, 8 (2019); C. Develotte and E. Rechniewski, "Discourse Analysis of Newspaper Headlines: A Methodological Framework for Research into National Representations," *Web Journal of French Media Studies* 4, no 1. (2001); T. Fogec, "Critical Discourse Analysis of Tabloid Headlines" (diploma thesis, Filozofski fakultet u Zagrebu, 2014), http://darhiv.ffzg.unizg.hr/id/eprint/5215/; N. Aqromi, "An Analysis of Metaphor for Corona on Headlines News," *Pioneer: Journal of Language and Literature* 12, no. 2 (2020): 157, https://doi.org/10.36841/pioneer.v12i2.734; M. Arshad and 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, no. 1 (2021): 176–86, https://doi.org/10.47264/idea.jhsms/2.1.15; D. Dor, "On Newspaper Headlines as Relevance Optimizers," *Journal of Pragmatics* 35, no. 5 (2003): 695–721, https://doi.org/10.1016/S0378-2166(02)00134-0.

[5] Anni Järvelin et al., "Information Retrieval from Historical Newspaper Collections in Highly Inflectional Languages: A Query Expansion Approach," *Journal of the Association for Information Science and Technology* 67, no. 12 (2016): 2928–46, https://doi.org/10.1002/asi.23379; Sanna Kumpulainen and Elina Late, "Struggling with Digitized Historical Newspapers: Contextual Barriers to Information Interaction in History Research Activities," *Journal of the Association for Information Science and Technology* 73, no. 7 (2022): 1012–24, https://doi.org/10.1002/asi.24608.

[6] R. Saha, A. Mondal, and C. V. Jawahar, "Graphical Object Detection in Document Images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)* (The Institute of Electrical and Electronics Engineers, Inc., 2019), 51–58; X. Yi et al., "CNN Based Page Object Detection in Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (The Institute of Electrical and Electronics Engineers, 2017), 230–35, https://doi.org/10.1109/icdar.2017.46.

[7] B. C. G. Lee et al., "The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (ACM, 2020), 3055–62, https://doi.org/10.1145/3340531.3412767.

[8] J. R. Chaudhary and J. Paulose, "Opinion Mining on Newspaper Headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering*, 9, no. 3 (2019): 2152–63, https://doi.org/10.11591/ijece.v9i3.pp2152-2163.