Leveraging NLP and AI to Enhance Quality Assurance in Education : A Case Study from Thailand

Punyisa Phumiphol

Office for National Education Standards and Quality Assessment (ONESQA)

Abstract

This paper presents a case study on the application of Natural Language Processing (NLP) and Artificial Intelligence (AI) to improve the benchmarking and decision-making processes in Thailand's External Quality Assurance (EQA) system. This study attempted to address this issue by creating an NLP-based data extraction pipeline adapted to EQA benchmarking requirements through the mix-methods research. Stakeholder surveys found that approximately 87.5% of respondents identified the need for automated NLP techniques for transforming unstructured data into actionable insights, implying that the response emphasizes the practical relevance of creating technology to expedite and improve the benchmarking process in education. Leveraging these findings, the NLP pipeline was built using regular expression, pattern matching, and Named Entity Recognition (NER) to capture the desired text from complicated documents. Thereafter utilizing TF-IDF to vectorize and analyze meaningful insights with high accuracy, reaching a 98.33% match with annotated datasets and an F1 score of 1.0, the system effectively extract data while also obtaining critical data to support advanced analytics and visualizations revealed hidden performance patterns for both regulatory and collaborative benchmarks.

Keywords: NLP, Data Extraction, Information Extraction, Quality Assurance

1. Introduction

The quality assurance (QA) was utilized as mechanism for optimal quality of education worldwide. The Office for National Education Standards and Quality Assessment (ONESQA) is a primary agency responsible in this part, this holds schools accountable to stakeholders and fosters transparency in educational procedures, which aids in benchmarking to established standards and best practices.

Considering the diversity of school contexts, and enormous volumes of qualitative data generates substantial hurdles (ONESQA, 2021). Furthermore, traditional analysis and visualizations frequently fail to show shortcomings and opportunities for growth.

Dependence on unstructured data presents substantial obstacles for information retrieval and analysis, manual data extraction from school annual reports poses several obstacles (ONESQA, 2024). According to these limits, there is an urgent need to improve quality evaluation systems to properly handle these challenges (Figure 1). This study aims to leverage natural language processing (NLP) techniques to extract data from school documents and evaluate performance against quality standards to inform the creation of meaningful benchmarks, thereby encouraging continuous improvement and development among schools.

1.1 Aims and Objectives

- To develop an NLP-based automated data extraction pipeline for EQA artifacts in Thai
- To demonstrate the extracted data utility through cluster analysis and visualization for EQA benchmarking

2. Literature Review: Use cases of NLP

2.1 Information Extraction

In practice, Malashin et al. (2024) used Optical Character Recognition (OCR) in combination

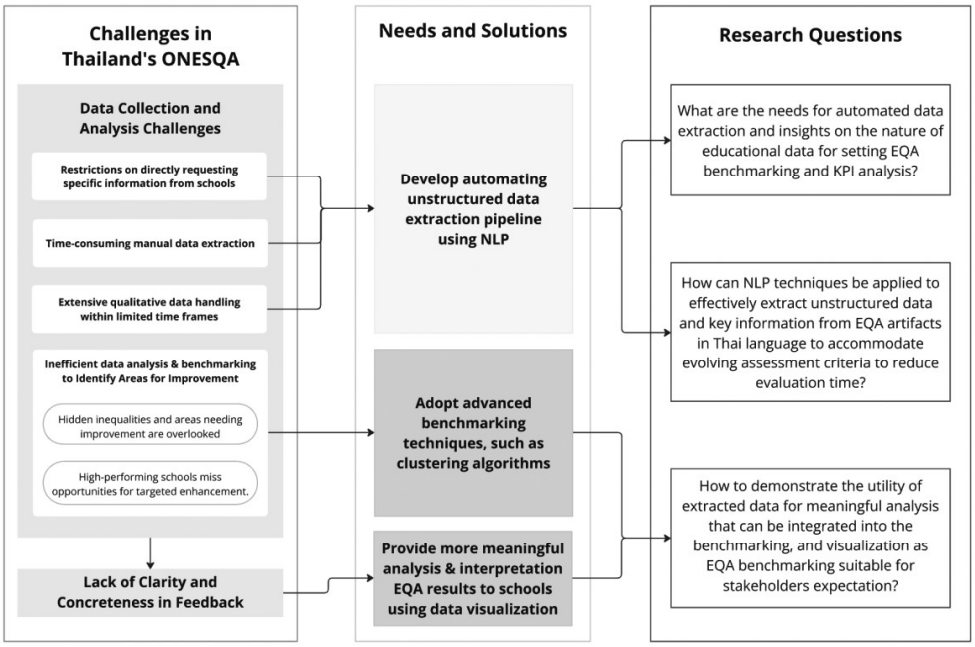

Figure 1: Research Gap

with Genetic Algorithms and Neural Networks to extract text and important information from document images which utilized PyTesseract and easyOCR for text recognition and Genetic Algorithms to improve OCR settings (Malashin et al., 2024). While another study by Hansen et al. (2019) focused on extracting and categorizing unstructured data from PDFs using OCR paired with deep learning techniques such as Faster Region-based Convolutional Neural Network (R-CNN), which aids in document segmentation and detection (Hansen, M., et al, 2019).

Overall, the case studies highlight the effectiveness of information extraction to automate the extraction of data from PDFs, OCR for transforming scanned documents into machine-readable text could be used. Regex are frequently used for finding and recognizing particular patterns in text. Referring to the approaches mentioned, a technique or method is unable to establish a solid data collection pipeline, considering select appropriate methods and creating combination of methods for specific goals is crucial to enabling the correct extraction.

2.2 Handling Specific Language Challenges

In practices, Soisoonthorn et al. (2023) demonstrated the actual implementation of these approaches by using SDR-based algorithms for Thai word segmentation and found the significant improvements in accuracy for languages, especially segmentation issues due to the lack of spaces between words (Soisoonthorn, T. et al., 2023). Correspondingly, Meesad (2021) applied Long Short-Term Memory (LSTM) networks to sequence handling in the context of fake news detection, proving the model's capacity to preserve context while improving classification accuracy (Meesad, P. 2021).

In addition to these techniques, recent research by Phatthiyaphaibun et al. (2023) created a set of tools to handle the Thai language's particular obstacles by integrating Conditional Random Fields (CRFs), LSTM, and Bidirectional Encoder Representations from Transformers (BERT) to solve tokenization and part-of-speech tagging constraints. The results of this architecture demonstrated that CRFs can be utilized for sequence labeling due to the lack of gaps between each word. While LSTM networks enable to capture long-term context and grasp the complete meaning of complicated statements, and BERT has strong capabilities for identify actual meaning relevant to the context surrounding that boost the accuracy of text classification in Thai (Phatthiyaphaibun et al., 2023).

However, developing NLP for Thai presents considerable hurdles because the structure of Thai language complexity, which includes different level tone of voice certain word can use to conveying soften a statement or command, making the language subtler and more nuanced phrases, a distinct set of Thai numerals, and lack of annotated dataset. As previously stated, Phatthiyaphaibun et al. (2023) suggest that future research should

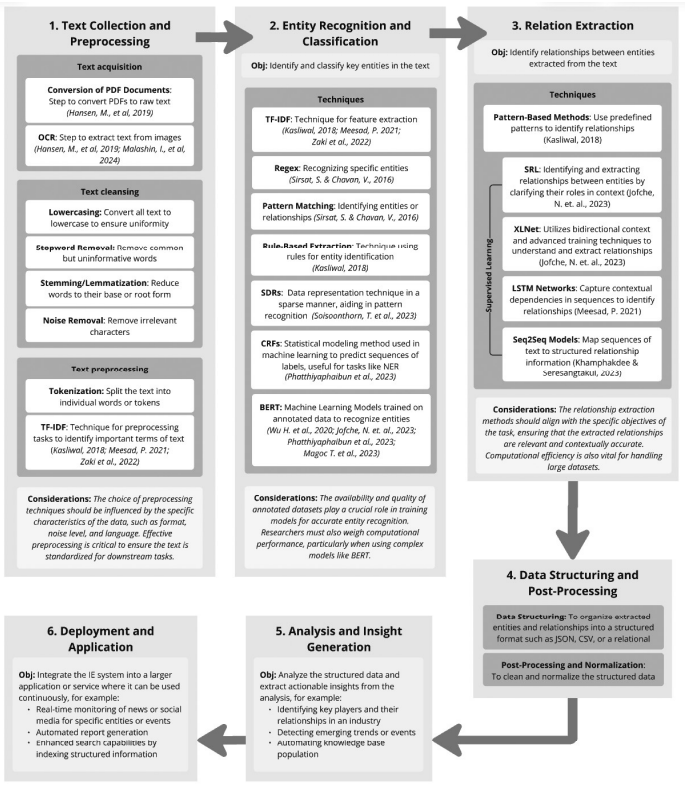

Figure 2: Summary of the Use Cases NLP Techniques for Data Extraction

concentrate on extending and diversifying datasets, as well as refining algorithms to more effectively handle Thai's distinctive linguistic qualities.

To conclude, Figure 2 illustrates the overview of structured framework for information extraction (IE), emphasizing the application of various NLP techniques and methodologies. The structured approach provided in this framework suggests a clear knowledge that each stage is planned to be executed with efficiency and in accordance with suitable principles.

3. Research Methodology

3.1 Research Methodology

This study adopts a mixed-methods approach to addressing the complexities of improving EQA benchmarking by automating data extraction using Natural Language Processing (NLP) to extract adequate information for advanced analysis. Data will be collected from two main sources including surveys of ONESQA leaders and officers and EQA artifacts.

3.2 Data Collection Methods

There are two phases of data collection methods from two primary sources including

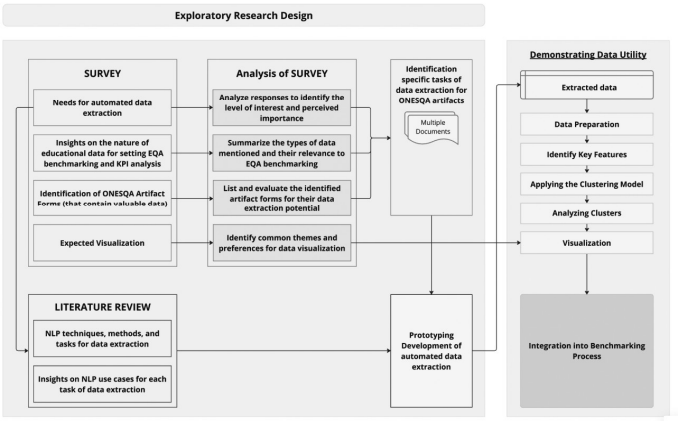

Figure 3: Research Design

surveys and EQA artifacts. Surveys administered to ONESQA internal and external stakeholders aim to identify the needs for automated data extraction, understand the nature of educational data for EQA benchmarking and KPI analysis (Figure 3). Subsequently, EQA artifacts, including reports and evaluations, will be processed using NLP techniques to extract both qualitative and quantitative data.

3.3 NLP-Based Automated Data Extraction Pipeline Framework

The developing NLP for EQA data extracting framework was designed to provide a clear

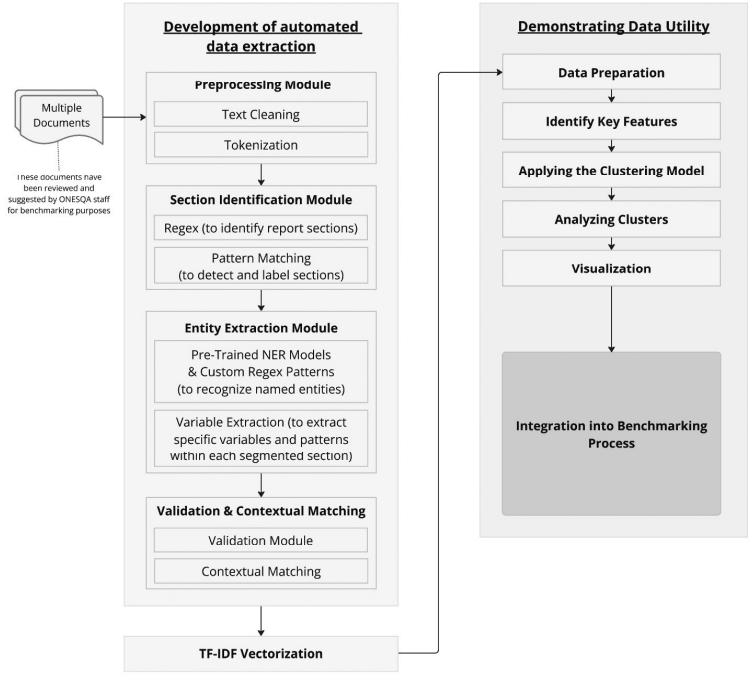

Figure 4: NLP-Based Automated Data Extraction Pipeline Framework

path, assuring systematic handling of each stage of data processing (Figure 4). The framework is segregated into particular tasks to improve modularity and simplify administration and debugging.

3.4 Data Analysis

The analysis is conducted in two phases, applying the following approaches.

1) **Survey Analysis**: Analyze using descriptive statistics and content analysis.
2) **EQA Artifact Analysis**: Develop and test an automated data extraction pipeline using NLP techniques. The extracted data will be statistically examined and compared to a manually annotated dataset using performance metrics of precision and recall measures.

4. Findings and Discussion

4.1 Needs for automated data extraction

According to the survey findings, 64.8% to 88.8% of respondents, including internal and external stakeholders, acknowledged the general necessity for an NLP data pipeline to extract adequate variables for comparison analysis (Figure 5). Specifically, 87.5% to 88.8% of respondents stated a need for developing NLP pipeline.

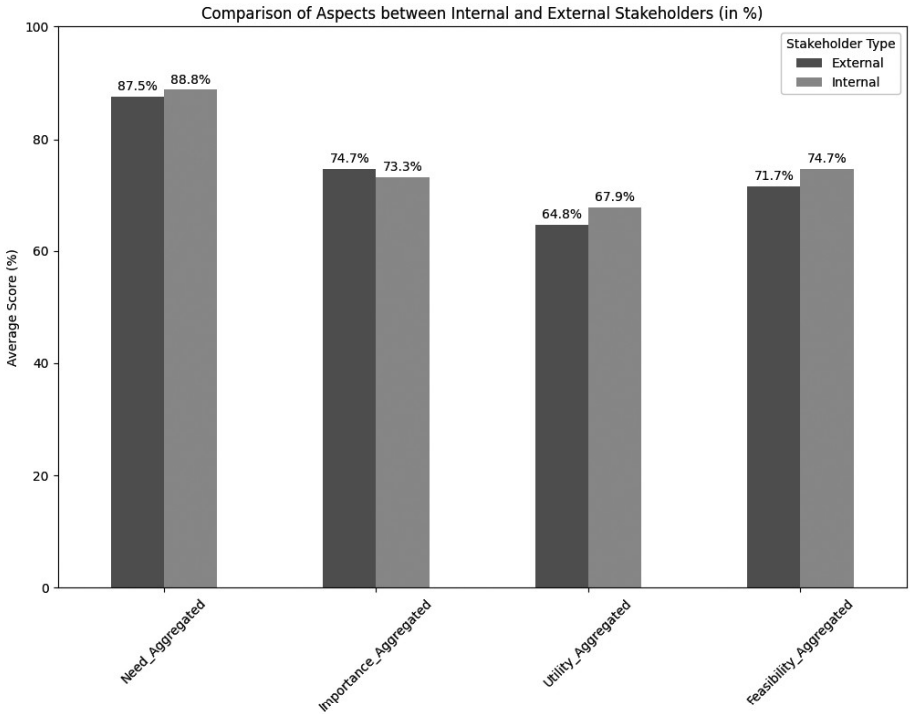

Figure 5: Comparison of needs aspects between internal and external stakeholders

4.2 Development and Performance of the NLP Extraction Pipeline

4.2.1 NLP Extraction Pipeline Development

1) Analysis of "BF-02" Document

The "BF-02" document suggested by respondents contains all essential variables (Figure 6). It has numerous organized portions that regularly address both Early Childhood Education (ECE) and Basic Education (BE). Each section presents indicators with corresponding ratings (numerical scores) and best practice descriptions (textual content).

Key challenges encountered during the analysis included repetitive headings that had identical headings, necessitating differentiation in the code to avoid confusion between sections. Due to inconsistent formatting while most portions maintained a consistent style. These issues were overcome by meticulous code design and optimization, ensuring that the

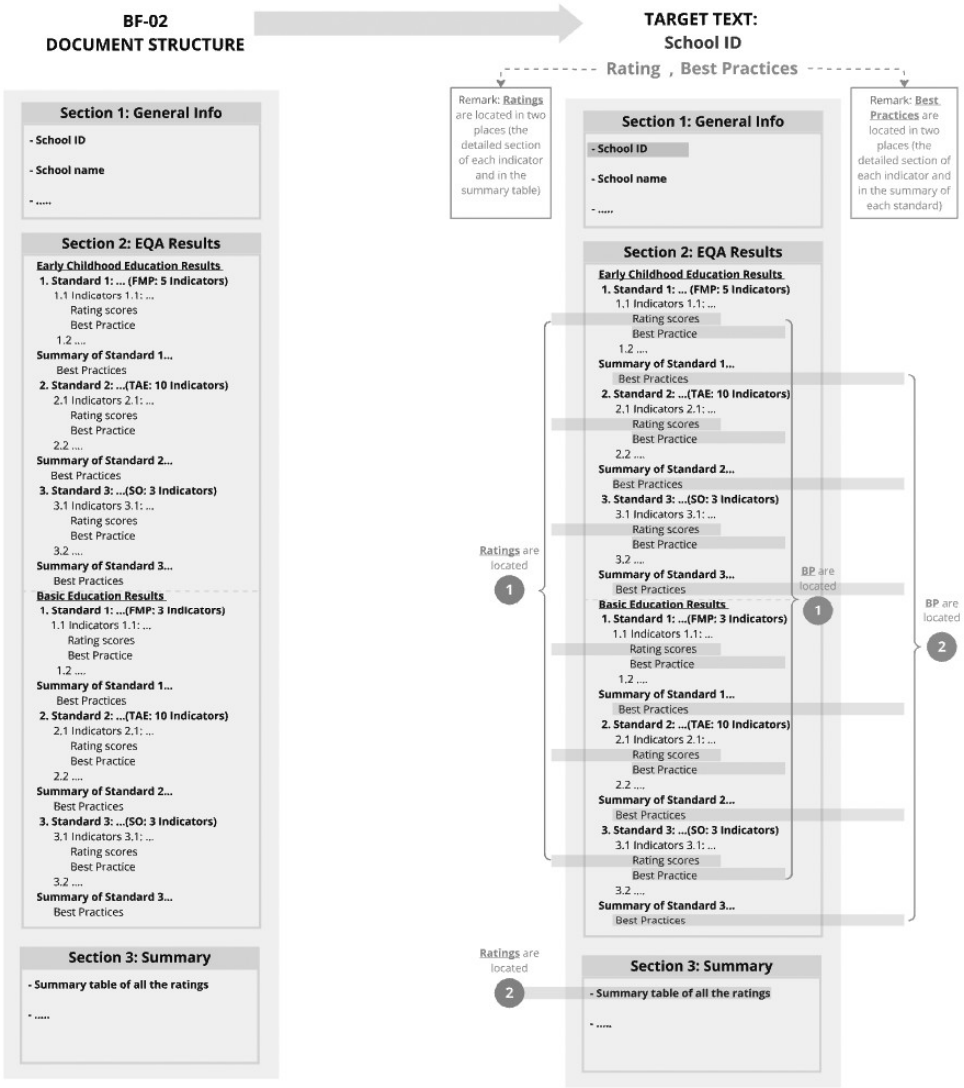

Figure 6: "BF-02" Document Structure

pipeline could successfully extract the required data.

2) Findings

2.1) Part 1: Rating Extraction

This developed code aims to extract numerical ratings for each indicator in both ECE and BE, which focuses on these precise places to effectively collect the numerical data associated with each indicator (Figure 7) focused on **School Code Extraction** contains a regex-based approach for extracting the school code. **Indicator and Rating Extraction** intended to obtain ratings for each indicator within a standard by using regex to identify each indicator by number to retrieve the related rating.

2.2) Part 2: Best Practices Text Extraction

This model is designed to extract the descriptive text under the "Best Practice" sections for each indicator. This part deals with extracting narrative text, requiring different NLP techniques to identify and isolate the relevant content ensuring that no important

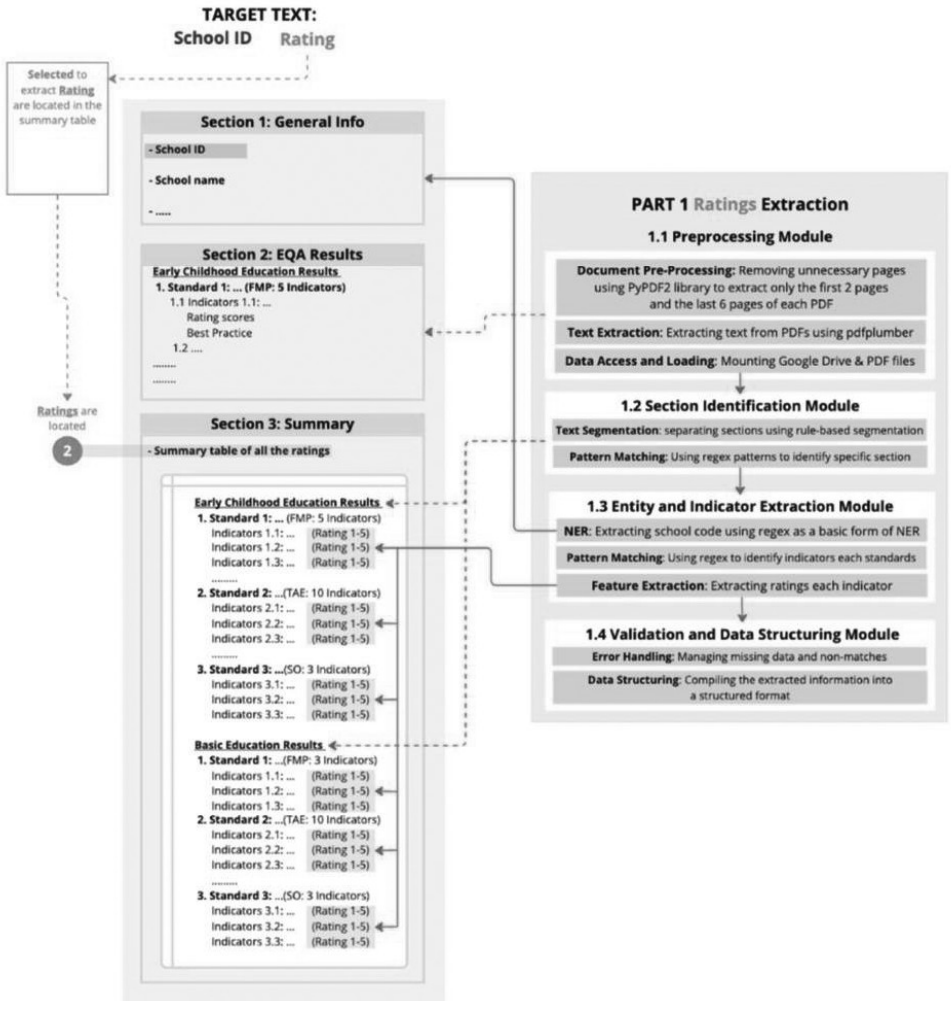

Figure 7: The Structure of Rating Extraction Development (Code 1)

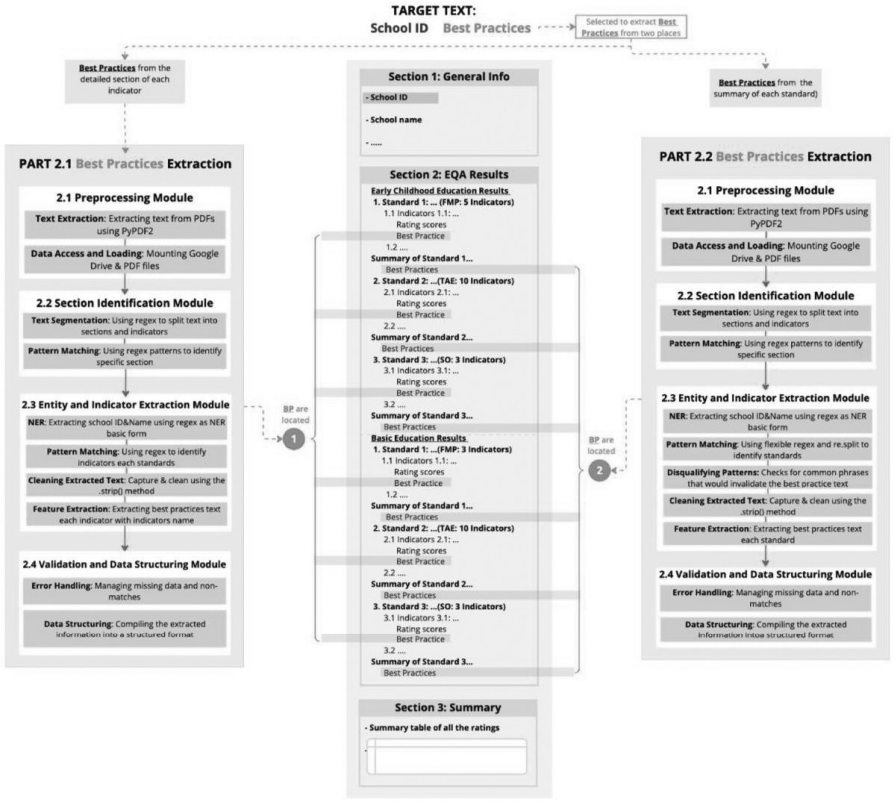

Figure 8: The Structure of Best Practices Extraction Development (Code 2.1-2.2)

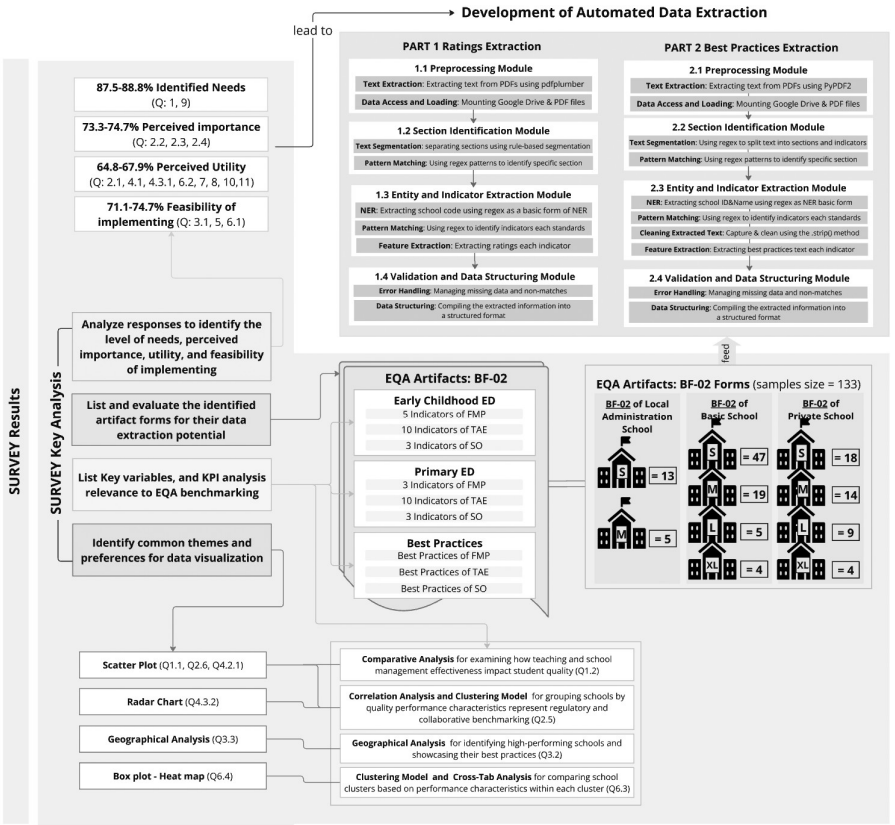

Figure 9: The results of NLP Extraction Pipeline Development

information was missed. These two codes focus on retrieving certain bits of data such as school codes, indications, and ratings using NER combined with regex and **Pattern Matching for Indicators** to capture each indicator's number and name. Then **Feature Extraction for Ratings**, the code identifies and extracts the rating or score associated with that particular indicator.

4.2.2 Evaluation Metrics for NLP Extraction Performance

1) Rating of Indicators Extraction Performance

The NLP extraction technique captures indicator ratings with high efficiency indicating by highest scores all performance metrics (Figure 10). Precision scores were perfect at 1.00, indicating that the model continuously returned correct ratings. Similarly, 1.00 recall values indicate that the pipeline successfully obtained all relevant ratings across indications, and 1.00 on the F1 Score confirms the balance between precision and recall metrics implying the overall reliability and robustness of the developed model.

2) Best Practices Text Extraction Performance

The NLP extraction model's performance in identifying relevant best practices from the BF-02 document was evaluated using standard metrics. Figure 11 shows consistently strong scores across all parameters, with accuracy approaching 1.0 (1.00 precision, 0.96 recall, and

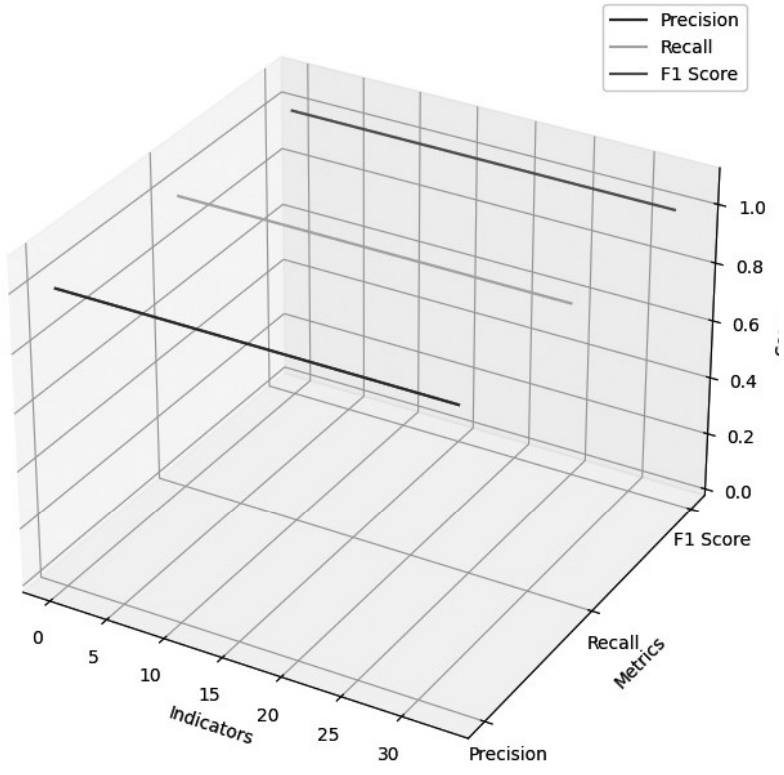

Figure 10: Rating of Indicators Extraction Performance

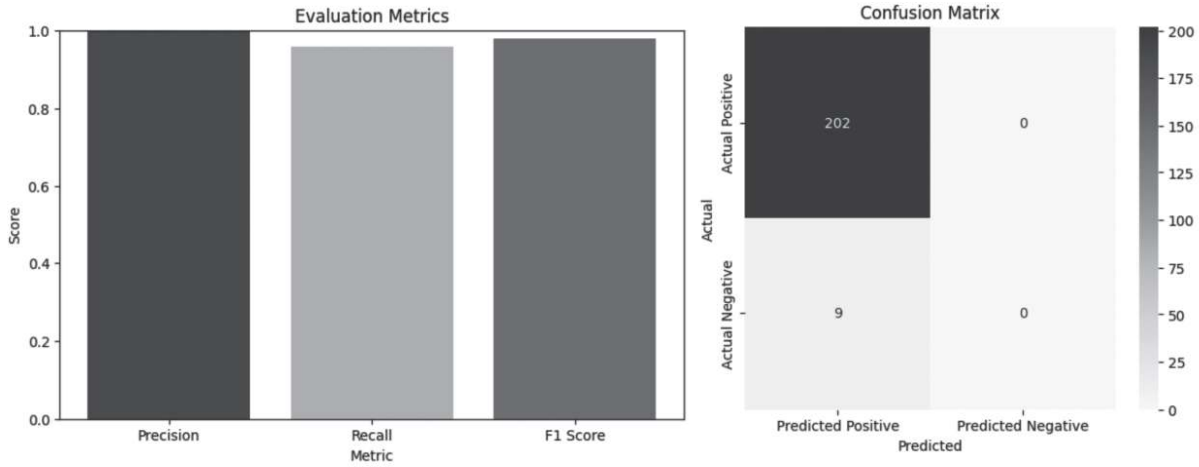

Figure 11: Best Practices Text Extraction Performance

0.98 F1 score).

To clarify the ability of this extraction, the cosine similarity was occupied to compare the two datasets. The results show that 236 out of 240 rows in the extracted dataset achieved high-quality matches, 98.33% of high-quality matches, and 1.00 average similarity score representing accuracy performance. Considering the confusion matrix, there are 202 identifiers from the annotated dataset were correctly matched with those in the extracted dataset, representing 81.21% of the total annotated identifiers.

Overall, these findings demonstrate the model's outstanding precision and recall, which resulted in practically alignment of the extracted and annotated data. Despite the small differences, the results demonstrate the model's ability to identify text-relevant best practices in the documents. This means that the built pipeline can replace the manual extraction process.

4.3 Application of Extracted Data for EQA Benchmarking

The following sections examine and show how the data was analyzed and displayed in order to achieve the goal of advanced analysis and visualization, which lead to improve both **regulatory benchmarking** and **collaborative benchmarking** processes.

1) Regulatory Benchmarking

1.1) Quadrant-Based Clustering

To create clusters, median values of Facilities & Management Performance (FMP) and Teaching Efficiency (TAE) were used to divide the schools into distinct groups instead of mean value because of the abnormal distribution. The results of this method helped manually identify groups of schools with similar performance characteristics, laying the groundwork for developing acceptable EQA regulatory standards based on their strengths

Table 1: School Performance Clustering Results Using Quadrant-Based

Cluster	Name	Teaching Efficiency (TAE)	Facilities & Management Performance (FMP)	Student Outcomes (SO)
Cluster 1	High Performers	Higher the median	Higher the median	Excellent
Cluster 2	Teaching-Focused Achievers	Higher the median	Lower the median	Very good to Excellent
Cluster 3	Teaching-Oriented Schools	Higher the median	Lower the median	Good to Adequate
Cluster 4	Well-Managed Potential	Lower the median	Higher the median	Very good to Excellent
Cluster 5	Facility-Driven Schools	Lower the median	Higher the median	Good to Adequate
Cluster 6	Low Performers	Lower the median	Lower the median	Good to Adequate
Cluster 7	Mixed Performers	Mixed	Mixed	Mixed

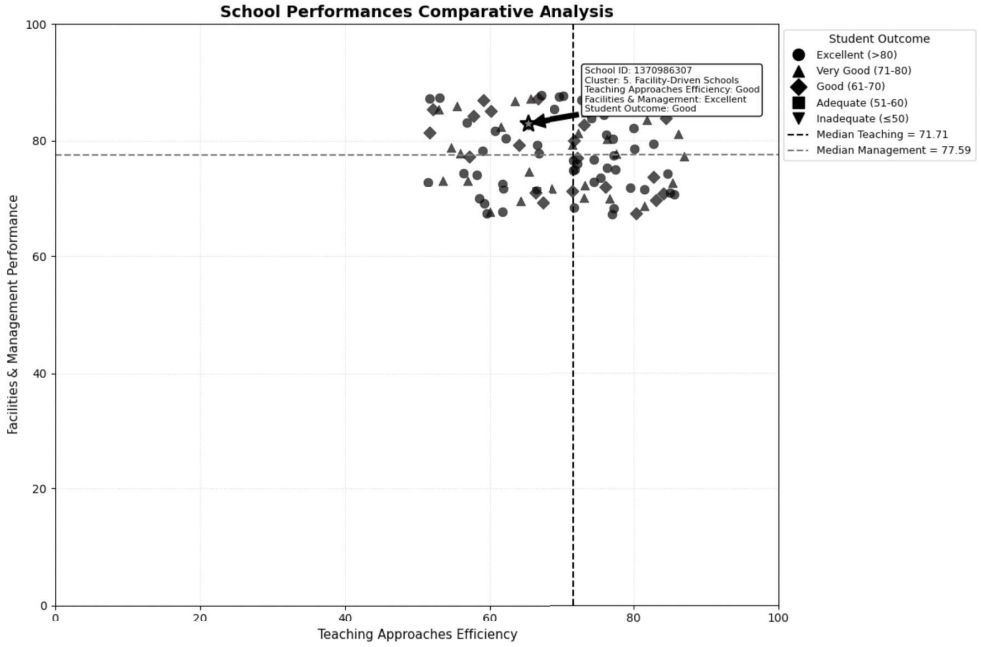

Figure 12: Scatter Plot of Comparative Analysis with Quadrant-Based Clustering

and shortcomings (Figure 12).

1.2) K-Means Clustering for Mixed Performers

These sub-clusters deliver additional specificity for analyzing school performance and recommend more targeted actions. **High performers**, for example, may provide a good example for others, nevertheless **mixed performers** would benefit from tailored tactics that addressed particular areas of weakness (see Figure 13 and Table 2). By categorizing schools, figures 14 and 15 show how these clustering algorithms benefit EQA by recognizing strengths and weaknesses across schools.

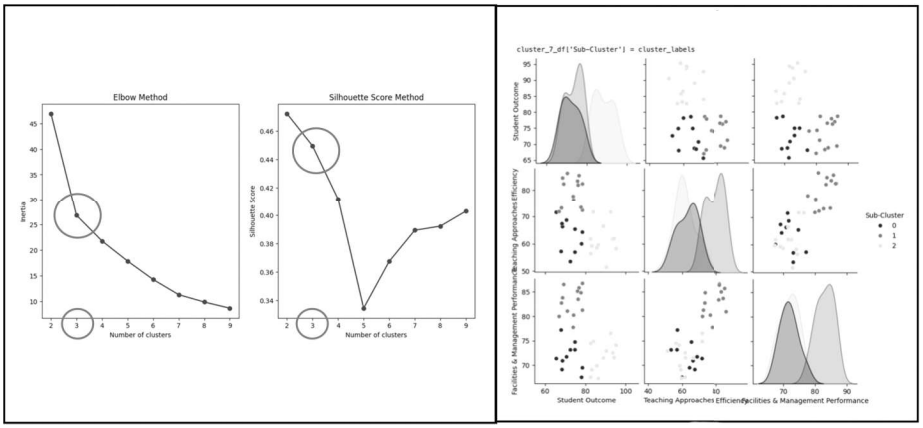

Figure 13: K-Means Clustering of Sub-Group

Table 2: The Summary of School Performance Clustering Results Using Quadrant-Based and K-Means Algorithms

Cluster	TAE	FMP	SO	Clustering Reason
Cluster 7: Outcome-Focused Achievers	Close to the median both above and lower, near the line of median	Close to the median both above and lower, near the line of median	Very good to Excellent (green and dark green dots)	Schools have high SO while performing averagely in other areas, FMP and TAE.
Cluster 8: Balanced Performers	Higher the median	Higher the median	Close to the median (light green and green dots)	Schools excel in FMP and TAE but have average SO, indicating a balanced focus on resources and teaching, with room for improvement in outcomes
Cluster 9: Consistent Moderates	Close to the median both above and lower, nearly the median line	Close to the median both above and lower, nearly the median line	Close to the median (light green and green dots)	Schools have moderate performance across all variables, without any particular area standing out as either strong or weak.

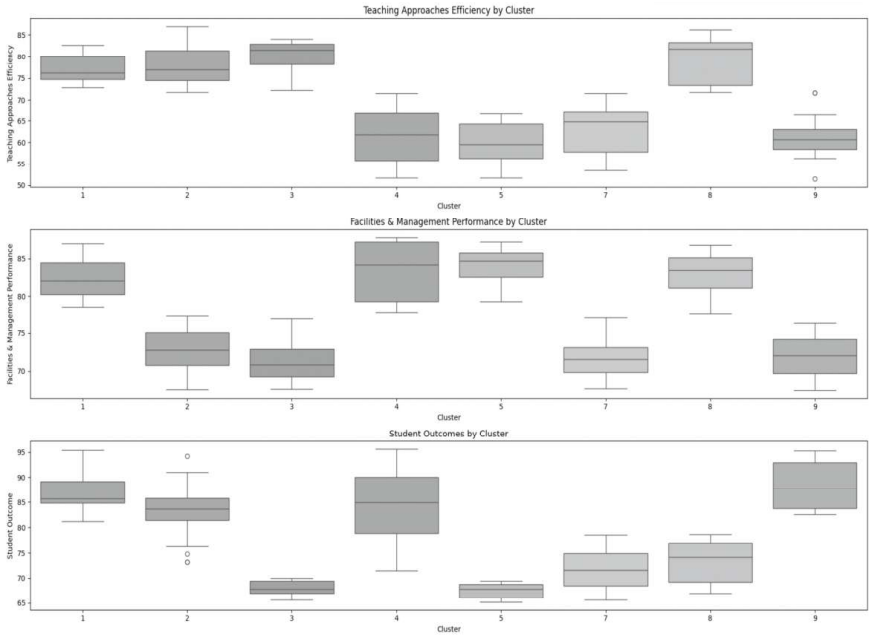

Figure 14: Box Plot of School Performance Characteristics by Cluster

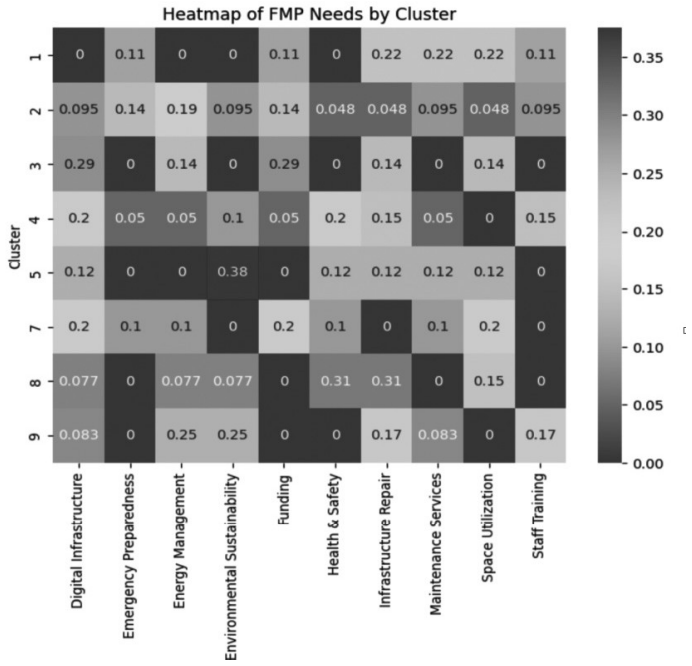

Figure 15: Heat map of Performance Needs Across School Clusters

2) Collaborative Benchmarking

2.1) Best Practices and Geographic Analysis

This visualization considers that when a school needs to adopt a best practice, they should first consider the similar demographic before selecting a project that could work for their situation so that demographic data will be taken part in identifying high-performing schools and their best practice alongside the goal of facilitating adoption by other schools (Figure 16 and 17).

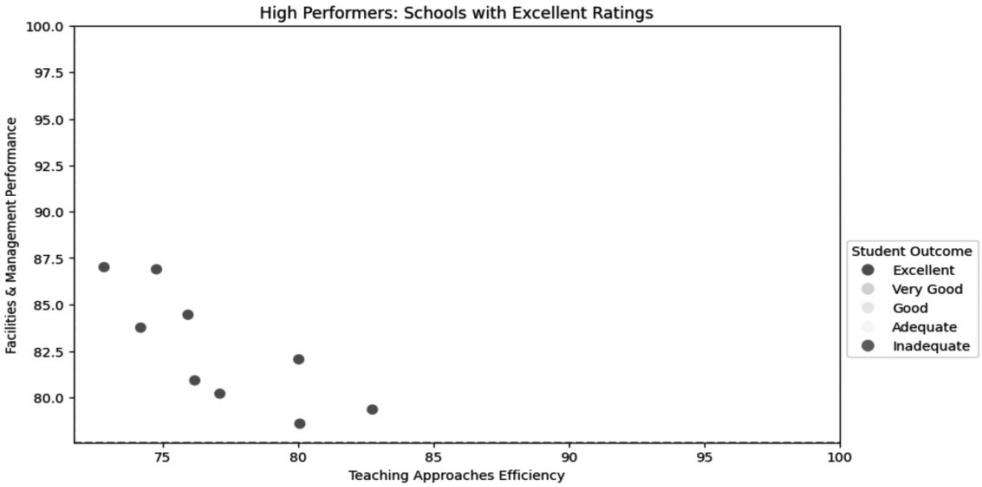

Figure 16: Scatter Plot of High-Performing Schools with Excellent Ratings

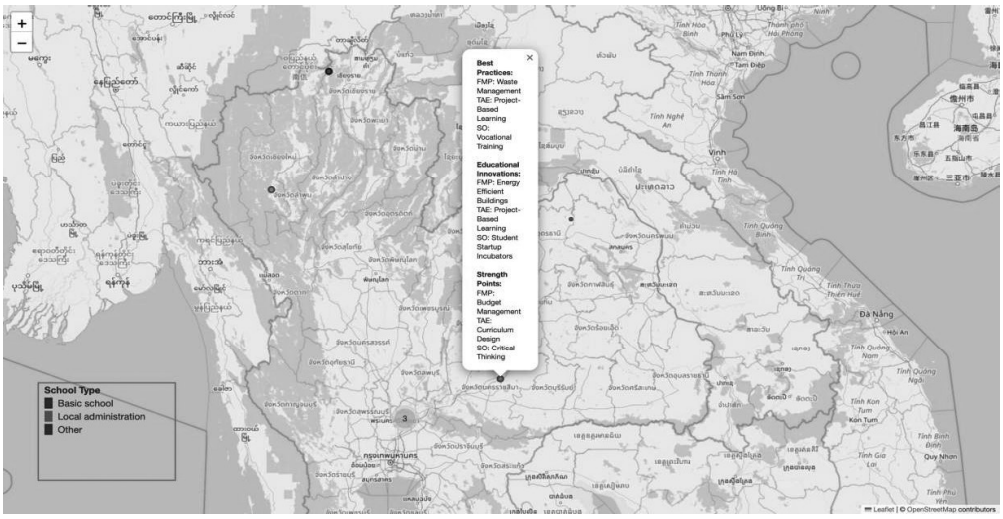

Figure 17: Geographic Map of High-Performing Schools with Best Practices for Adoption

2.2) Individual School Performance and Benchmarks

Using data visualizations as individual school performance scoreboards (Figure 18), to compare schools with established benchmarks, both overall and within their clusters. Moreover, the visualization empowers schools to set their own goals, whether they aim to meet average performance levels or continue excelling beyond their current benchmarks compared with cluster benchmarks, high-performer benchmarks, and overall benchmarks (Figure 19).

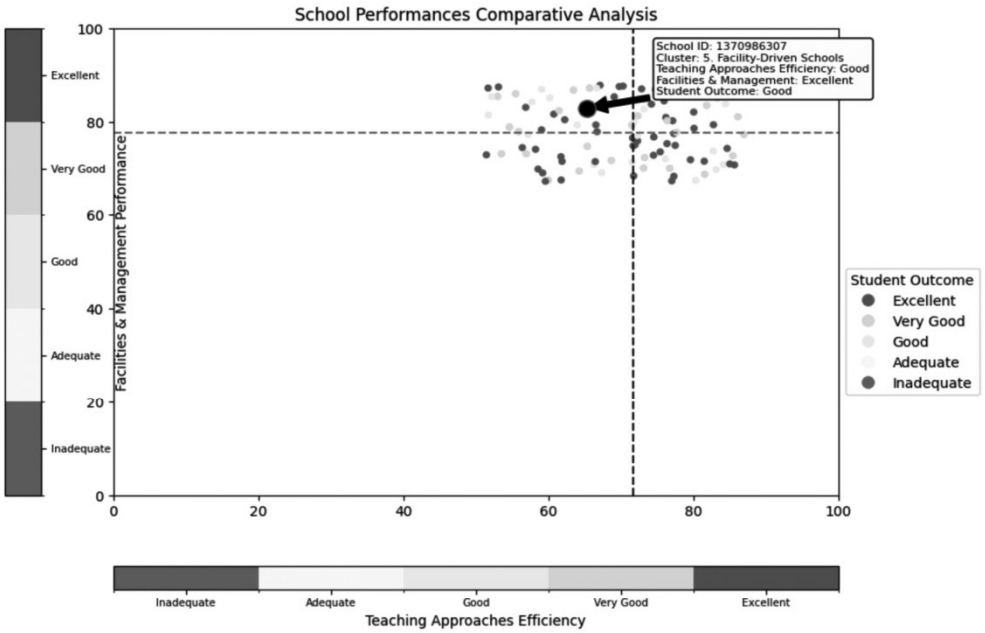

Figure 18: Scatter Plot of Individual School Performance with Ranking Comparison

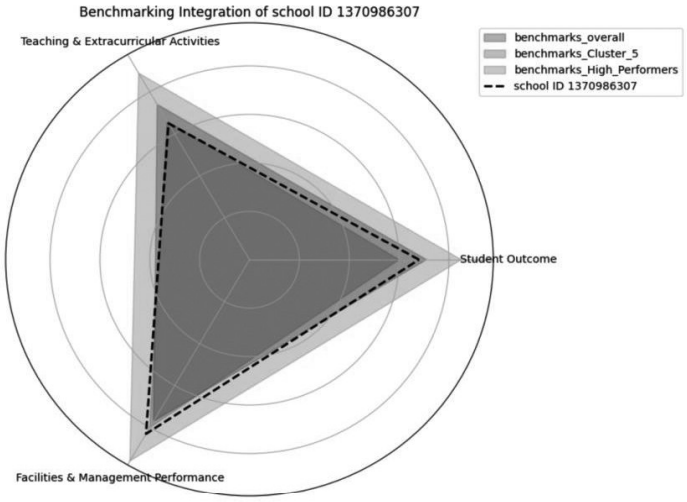

Figure 19: Radar Chart of Individual School Performance

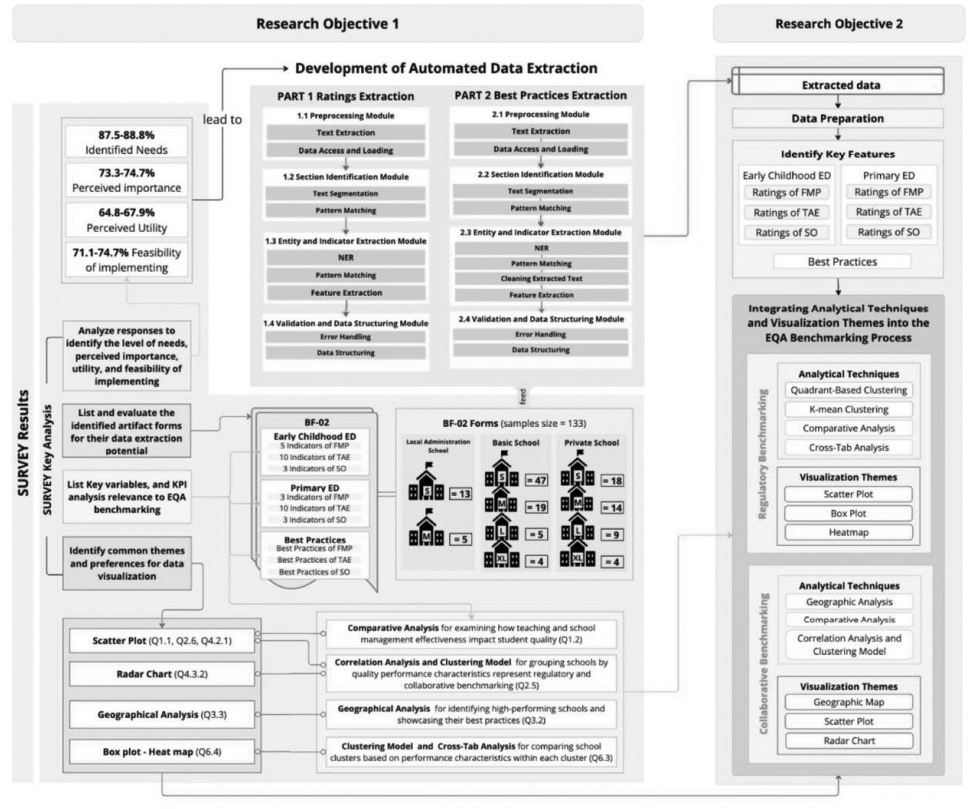

Figure 20: Research Results

4.4 Discussion

4.4.1 Key Insights of NLP Data Extraction Development

According to Zaki et al. (2022), preprocessing and vectorization improve data dependability, as does the NLP-based extraction pipeline, which extracts both numerical ratings and descriptive best practices accurately (98.33%). In addition, Chen et al. (2023), who highlight

the role of regex and text segmentation in improving extraction accuracy. However, despite its success with a small dataset of 133 samples, the model must be adapted to handle the larger datasets typical of the EQA process, often exceeding 1,000 files. To ensure accurate extraction on this scale, the algorithm requires enhancements in table recognition to capture ratings embedded in tables effectively.

A significant challenge emerged in the segmentation of Thai text, where the absence of explicit word boundaries reduced the pipeline's recall rate, which was obvious in the 15.27% of identifiers (38 out of 249) that were neither successfully matched nor marked as missing. This is parallel to the findings of Soisoonthorn et al. (2023) and Phatthiyaphaibun et al. (2023), who investigated Thai language-specific restrictions in NLP. To overcome this obstacle, future research should focus on employing language-specific models to effectively solve segmentation problems and improve overall accuracy.

4.4.2 Application of Extracted Data for EQA Benchmarking

The key findings demonstrate the utility of extracted data for advanced analysis and visualizations, which is a sufficient input for EQA benchmarking that evidently reveals hidden patterns of school performance characteristics that cover regulatory and collaborative benchmarking performance across critical areas. These tools are more effective than traditional descriptive statistics or correlation analysis, which often fail to reveal hidden patterns and non-linear relationships in the data (Jo, 2023).

The incorporation of extracted data into the benchmarking process is effective and adaptive, which leads to dynamic, context-sensitive benchmarks. This confirms the views of Lucander & Christenson (2020), and Marciniak (2018), who argue that when executed correctly, benchmarking allows institutions to monitor performance against standards and pursue continual improvement. The point is that this proposes analysis and visualizations facilitating benchmarking by giving a macro (systemic) and micro (individual school) awareness of strengths and deficiencies. This dual perspective aligns with Tangpornpaiboon's (2022) critique, as it allows ONESQA to deliver more personalized and actionable recommendations, bridging the gap between data collection and meaningful feedback.

CONCLUSION

The NLP pipeline's efficiency was proved by its excellent accuracy rates, which included a 98.33% match with annotated datasets and an F1 score of 1.0 for relevant text of best practices and numerical ratings extraction. Understanding and implementing stakeholder demands enabled the pipeline to effectively acquire vital data with minimum differences,

hence supporting the objective of enhancing automated EQA methodologies.

The utility of extracted data by analytics and visualization proposed in this study enhances stakeholder engagement, transparency, and accountability by making the EQA process more data-driven and objective. Besides, for guiding benchmark setting and supporting schools on their improvement paths driving sustainable progress in education.

This study enhances the area of educational quality assurance by demonstrating how an NLP-based data extraction pipeline can accelerate ahead of the previously time-consuming benchmarking procedure. Future study ought to aim at developing these approaches and examining their broader applicability to establish systems that allow for continuous improvement and support the continued growth of educational quality assurance standards.

References

Chen, Qing, Angello Banerjee, Çagatay Demiralp, Greg Durrett, and Isil Dillig. "Data Extraction via Semantic Regular Expression Synthesis." *Proceedings of the ACM on Programming Languages* 7 (2023): 1848–77. https://acm.org.

Hansen, Malte, Andreas Pomp, Katharina Erki, and Tobias Meisen. "Data-Driven Recognition and Extraction of PDF Document Elements." *Technologies* 7, no. 3 (2019): 65. https://doi.org/10.3390/technologies7030065.

Jo, Ara. "Visualization And Analysis Of Student Data Using Machine Learning And Statistical Methods." Master's thesis, Tampere University, 2023. https://tuni.fi.

Jofche, Nikolche, Kire Mishev, Riste Stojanov, Milos Jovanovik, Eftim Zdravevski, and Dimitar Trajanov. "PharmKE: Knowledge Extraction Platform for Pharmaceutical Texts Using Transfer Learning." *Computers* 12, no. 1 (2023). https://doi.org/10.3390/computers12010017.

Kasliwal, Nikhil. Natural Language Processing with Python Quick Start Guide: Going from a Python Developer to an Effective Natural Language Processing Engineer. Birmingham: Packt Publishing, Limited, 2018. ProQuest Ebook Central.

Kayyali, Mohammad. "The Relationship between Rankings and Academic Quality." *International Journal of Management, Sciences, Innovation, and Technology* 4, no. 3 (2023): 1–11. https://www.researchgate.net/publication/371982309_The_Relationship_between_Rankings_and_Academic_Quality.

Khamphakdee, Nattasate, and Paskorn Seresangtakul. "An Efficient Deep Learning for Thai Sentiment Analysis." *Data* 8, no. 5 (2023): 90. https://doi.org/10.3390/data8050090.

Lucander, Helene, and Cecilia Christersson. "Engagement for Quality Development in Higher Education: A Process for Quality Assurance of Assessment." *Quality in Higher Education* 26, no. 2 (2020): 135–55. https://doi.org/10.1080/13538322.2020.1761008.

Magoc, Tanja, Reid Everson, and Christopher A. Harle. "Enhancing an Enterprise Data Warehouse for Research with Data Extracted Using Natural Language Processing." *Journal of Clinical and Translational Science* 7, no. 1 (2023). https://doi.org/10.1017/cts.2023.575.

Malashin, Ivan, Ivan Masich, Vladimir Tynchenko, Alexander Gantimurov, Vladimir Nelyub, and Anton Borodulin. "Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports." *Machine Learning and Knowledge Extraction* 6, no. 2 (2024): 1361–77. https://doi.org/10.3390/make6020064.

Marciniak, Rafał. "Quality Assurance for Online Higher Education Programmes: Design and Validation of an Integrative Assessment Model Applicable to Spanish Universities." *International Review of Research in Open and Distributed Learning* 19, no. 2 (2018): 126–54. https://doi.org/10.19173/irrodl.v19i2.3443.

Meesad, Phayung. "Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and

Machine Learning." *SN Computer Science* 2, no. 6 (2021): 425. https://doi.org/10.1007/s42979-021-00775-6.

Office for National Education Standards and Quality Assessment (ONESQA). "การศึกษารูปแบบการประเมินคุณภาพภายนอกแบบเสมือนจริง [Study on Virtual Models for External Quality Assessment]." จุลสาร สมศ. *[ONESQA Journal]* 22, no. 1 (2021): 6–12. https://www.onesqa.or.th/upload/download/202204051424371.pdf.

Office for National Education Standards and Quality Assessment (ONESQA). *ONESQA Self-Evaluation Report*. Internal report, ONESQA, 2024. https://www.onesqa.or.th/upload/download/202405091620022.pdf.

Phatthiyaphaibun, Wannaphong, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Peerat Chormai, Pattarawat Limkonchotiwat, Thanakorn Suntorntip, and Can Udomcharoenchaikit. *PyThaiNLP: Thai Natural Language Processing in Python*. Ithaca, NY: Cornell University, 2023. https://arxiv.org/pdf/2312.04649.pdf.

Sirsat, S. R., and V. Chavan. "Pattern Matching for Extraction of Core Contents from News Web Pages." *In 2016 Second International Conference on Web Research (ICWR)* , 2016. https://www.researchgate.net/publication/304456801_Pattern_matching_for_extraction_of_core_contents_from_news_web_pages.

Soisoonthorn, Teerayut, Helfried Unger, and Montri Maliyaem. "Thai Word Segmentation with a Brain-Inspired Sparse Distributed Representations Learning Memory." *Computational Intelligence and Neuroscience* 2023 (2023): 1–10. https://doi.org/10.1155/2023/8592214.

Tangpornpaiboon, Saruta. "Understanding Discrepancies in Trends Results of PISA, TIMSS, and O-NET and Implications for Education Policies in Thailand." PhD diss., University College London, 2021.

Wu, Hulin, Joseph M. Yamal, A. Yaseen, and Vahid Maroufy. Statistics and Machine Learning Methods for EHR Data: From Data Extraction to Data Analytics. Milton: CRC Press LLC, 2020. ProQuest Ebook Central.

Zaki, Nazar, Sherzod Turaev, Kadhim Shuaib, Anusha Baby, and Eman Mohamed. "Automating the Mapping of Course Learning Outcomes to Program Learning Outcomes using Natural Language Processing for Accurate Educational Program Evaluation." Preprint, posted October 2022. https://doi.org/10.21203/rs.3.rs-2196467/v1.