This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-93429-2_18. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

Applied optical character recognition and large language models in augmenting manual business processes for data analytics in traditional small businesses with minimal digital adoption

Zachary Matthew Alabastro, Joseph Benjamin Ilagan, Lois Abigail To, and Jose Ramon Ilagan

Business Insights Laboratory for Development, Ateneo de Manila University, Quezon City, Philippines
zachary.alabastro@student.ateneo.edu,
{jbilagan,jrilagan,lto}@ateneo.edu,
WWW home page: https://www.ateneo.edu/

Abstract. The local business landscape in the Philippines presents a slow adoption to digitization. There are fears of inaccurate results and limited technical proficiency. Data analytics and artificial intelligence (AI) are crucial for business growth. By understanding the performance of stock-keeping-units (SKU) with technology, businesses can find better ways to handle products to increase profitability. In exploring optical character recognition (OCR) and large language models (LLM), a software pipeline can help businesses analyze handwritten sales data for business intelligence. The research aims to develop a system for translating logbooks into metrics by augmenting manual business processes performed by staff. The application combines Amazon Web Services' OCR technology with Anthropic's Haiku 3.0 LLM. The pipeline extracts handwritten text from images and performs few-shot learning-based classification. Data from a local food stall was used for testing. Precision and recall scores were calculated to analyze similarities between original and extracted data. Results showed moderate-low precision but above-average recall. The SKUs yielded average precision and recall scores of 0.32 and 0.62, while sales data stood at 0.55 and 0.54, respectively. These scores indicate that the application struggles with accuracy but captures a fair amount of true values. Improvements are needed to enhance the learning mechanism of the LLM. Despite this, the application holds promise for helping small businesses achieve digitization to supplement manual business processes, taking into account the limited technological literacy of staff. It serves as a potential catalyst for growth by simplifying complex data problems with cost-efficient solutions.

Keywords: Optical character recognition · artificial intelligence · large language models · Python · data engineering · AI · prompt engineering · food stalls · local enterprises · sales analytics · staff augmentation

1 Introduction

The local business context in some Global South countries like the Philippines exhibits a slow adoption process to digital technology, especially for smaller-sized enterprises. Businesses are hesitant in recognizing the benefits of digitization due to fear of inaccuracies and learning curves [13]. Smaller organizations tend to be more risk-averse, avoiding potential change and uncertainty [12]. This behavior comes from a general misunderstanding of factors affecting such changes and the context surrounding the shift. For instance, artificial intelligence (AI) is perceived to replace human jobs, which affects adoption. People deem it as a threat that is powered to lower costs of operations and production [11]. Although, some small firms simply lack the proper resources to pursue such innovations without detrimentally impacting the scalability of their own business [21]. However, AI can be seen to augment business processes when applied with traditional staff activities through co-piloting strategies [9]. By applying ethical considerations in AI, business development can excel without replacing human jobs [17]. Rather, AI processes are supported by human intervention to produce maximum efficiency.

Data analytics is an essential element in business optimization [4]. Deep and relevant data allow the organization to identify actionable insights which can innovate the way they handle core activities while maintaining existing processes [13]. In the scope of sales data, adequate collection of such information can assist business owners in locating the high- and low-performance items stock-keeping-units (SKU), which offers solutions on inventory improvement, cost reallocation, category assortment strategies, and automated systems for sales data transformation [4,15,28]. AI has also been observed to complement operations management by identifying points for optimization in supply chain management, scheduling, and holding inventory [14].

Although raw sales data initially come as numbers from manual staff logging, data engineering can develop such figures into more digestible features that can tell a story to its audience. An advancement in machine learning (ML) is the optical character recognition (OCR) on images, featuring different attributes that can be useful for turning sets of handwritten logbook information into a database through extraction and document classification [6,20]. Large language models (LLMs) could perform few-shot learning-based classifications by identifying similar text that correspond to a specific item from a small amount of training information [22]. This retrieves power from its natural language processing (NLP) capabilities and prompt engineering [7]. Competitive advantage and innovation are determinants that affect business intelligence adoption. These are essential in determining the viability of the application [26].

1.1 Objectives and Research Questions

The objective is to explore how OCR and LLM technology can enable traditionally non-digital businesses such as small food stalls to optimize business activities without removing manual business processes for staff with limited proficiency in technology [9]. This can help businesses understand the written data from their

sales logbooks and translate them into understandable metrics. Although this is a feasible approach, there remains a concern with the accuracy of such a pipeline. The research questions ask:

RQ1) What are some considerations when developing the pipeline for the application?

RQ2) How accurate can a cost-efficient application be in order to produce a system with a low barrier to accessibility?

2 Review of Related Literature

The adoption of new information technology (IT) systems in small businesses proves vital for proper management support [13,23]. This can promote their competitive advantage within their specific business environment by leveraging automation of transactional tasks to divert focus on data-driven, strategic planning and direction. While there are a number of complexities within the sphere of IT adoption in smaller enterprises, this remains a potential factor that can motivate business growth [23]. However, caveats remain present in the adoption of technology for business processes. On one end, early adoption may allow for more innovation, but can impose risks due to poor networking, marketing and sales efforts. If the firm chooses to adopt later, it may provide insights on how such shift has impacted other similar businesses, but may also result in falling too far behind competitors. Hence, timing is a very essential factor to consider when leveraging the relevant technologies [29].

A study in Korea analyzed the accuracy of OCR technology in the context of a ship's logbook. Data from handwritten navigational information was captured and entered into a software application, yielding an error rate of approximately 11% [19]. Although limited in scope, this technological innovation can still be explored in other types of data, more particularly those with tabular formats.

Automated deep-learning, text extraction in logbooks are also being applied in other sectors like education. These methods are being used to reduce labor-intensive logging and provide valuable insight into attendance patterns while automating transactional workflows. With proper monitoring, the system holds a high degree of precision, which can streamline the process of logging data digitally [24]. Although in a more educational setting, such strategies can be used to keep employees more proactive towards heavier decision-making procedures [16].

AI-driven data pipelines is another tool that can increase the accuracy of predictive models. By integrating additional steps into the process, the accuracy of data extraction and transformation can be monitored. With newer ML technologies, this can be useful in automating manual data modification or correction, translating information into actionable insights quickly [10]. By allowing the machine to handle menial activities, especially when given larger amounts of data, the user could focus on deriving implicit information early on.

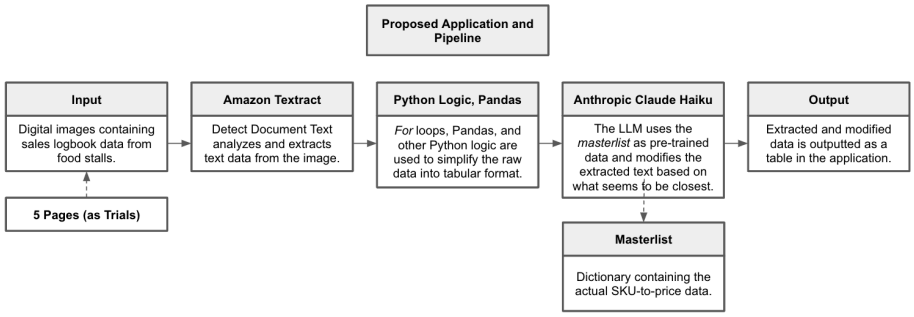


Fig. 1. Proposed application pipeline with OCR and LLM integration.

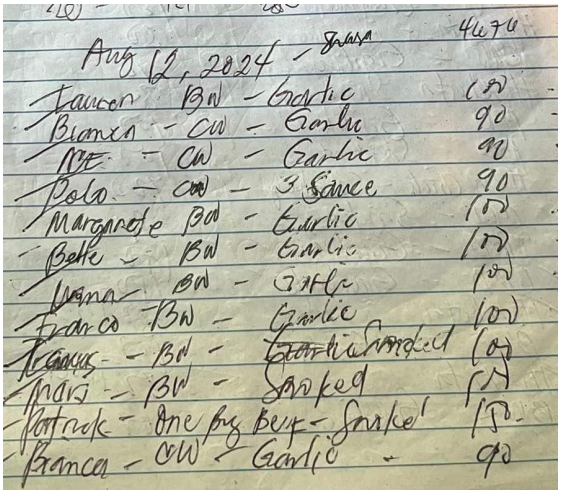


Fig. 2. Sample collection of sales data with handwritten logs from a food stall.

3 Methods

The application was built using Python, which integrated the OCR technology of Amazon Web Services (AWS) and the LLM application programming interface (API) of Anthropic Claude: Detect Document Text (DDT) and Haiku, respectively [3,2]. This was used as a means to provide cost-efficient alternatives while determining accuracy levels compared to other more expensive models. The pipeline consisted of the following steps as seen in Figure 1:

1. Manually upload the images of the sales logbooks digitally.
2. Extract the image with AWS OCR.
3. Access the handwritten blocks from the raw dictionary of results with Python logic.

4. Perform few-shot learning-based classification by letting the LLM read a *masterlist* of SKU-price data and retrieving information according to its perceived data content.

3.1 Data Collection

Data was collected from a local food stall in the Ateneo de Manila Student Enterprise Center to exhibit real-world results. Figure 2 shows an example of the data content, containing the name of the customer, an abbreviation of the SKU ordered, and the sales price per quantity ordered. It was used as a representative of other stalls in the same local cafeteria given the similarities of their respective sales log formats. The fast-paced environment of such stalls also make it so that abbreviated, handwritten logs are deemed as necessary to maintain their service during peak hours.

Five pages of logbook details were collected and uploaded into the software application. Upon extraction, the table of sales, their corresponding SKUs, and the confidence levels were presented.

4 Results

The results yielded the following similarity scores for precision and recall. These were obtained using a Python script with an analytical library, *scikit-learn* [27]. A sample result after extraction, including confidence level comments, can be seen Figure 3.

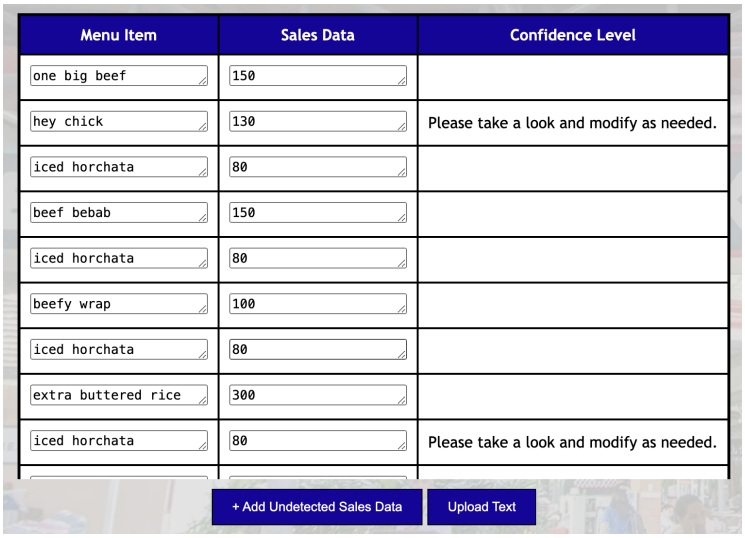


Fig. 3. Sample results of SKU-to-sales extraction data using the software application interface.

Table 1. Similarity scores for the SKUs between actual and predicted text.

	Page 1	Page 2	Page 3	Page 4	Page 5	Average
Precision	0.59	0.18	0.50	0.09	0.25	0.32
Recall	0.71	0.64	0.31	0.82	0.60	0.62

Table 2. Similarity scores for the sales between actual and predicted text.

	Page 1	Page 2	Page 3	Page 4	Page 5	Average
Precision	0.72	0.34	0.50	0.50	0.71	0.55
Recall	0.93	0.16	0.50	0.61	0.51	0.54

Average scores showed that the precision and recall for the extracted SKUs were at 0.32 and 0.62, respectively. The mean for precision and recall scores for sales data stood at 0.55 and 0.54, respectively. The prompt, "Please take a look and modify as needed," was displayed if AWS Textract DDT deemed its own extraction confidence level as below 85%.

5 Discussion

Results from the analysis yielded the following answers:

RQ1) What are some considerations when developing the pipeline for the application?

The pipeline revolves around handwritten data, with minor instances of printed formatting (i.e. content printed on the sheet prior to handwritten logging) to explore the viability of such technology towards the current contextual environment of the experiment. This means that manual transcription of orders are better suited to be processed by this application as it creates value and efficiency in post-work day analysis of SKU volume counts, orders achieved, and customer data, if any. It is also assumed that the data to be extracted has multiple lines of handwritten logs per day, in which the business has to manually translate into an online software for further analysis.

RQ2) How accurate can a cost-efficient application be in order to produce a system with a low barrier to accessibility?

There seems to be a neutral score among the five different pages, for both categories. With 1.00 being the desired score for both tools and 0.70-0.80 as the high-score threshold range, the current pipeline contributes a moderately-low precision rate, suggesting that the predicted values were unable to accurately output the true values [25,5,18]. However, in terms of recall scores, it seems to capture and output enough of the true values above 50%. Although the general pipeline contains more cost-efficient alternatives than other LLMs, the overall

accuracy must still be further refined. This application can work for smaller-scale businesses with an adequate amount of human intervention needed.

6 Conclusion

Although there were some outputs that produced the right results, there is still much to be analyzed about the integration of OCR and LLMs in the development of the software pipeline. This is essential to ensure that most of their long, menial tasks are automated and simplified. It serves as a tool for businesses and staff members to focus their attention on data-driven quality assurance (i.e. that of food products), strategic alignment, and direction. It also addresses the fear of job replacement and learning curves, as it proposes a simple way of managing data. By inputting an image into the system, the human receives a consolidated list of SKU-to-sales data with no complexities involved. It minimizes the proficiency levels required for adoption.

Small businesses require a catalyst for growth as part of the local and national economy. Entrepreneurs should be given the avenue to simplify complex data problems with cost-efficient and radical solutions, without sacrificing existing processes for its staff. As LLM and OCR technologies continuously innovate, the accuracies of the pipeline can also be improved.

6.1 Future Work

Prompt engineering is one aspect that can be improved on, as this also affects the learning mechanism of the model. By using other methods of training, the model can understand the context of such food products, allowing it to generate more accurate SKU and sales data. Fine-tuning is one approach to training LLMs with larger amounts of past data. This can improve the perception of the model after OCR extraction [8]. Further studies can also observe possible changes between the extraction results of AWS Textract DDT and the more advanced Analyze Document (AD). Although this may not align with the proposed cost-efficient pipelines, it may yield insights on how the overall software structure can be improved while maintaining minimal digital adoption [1].

Further iterations shall still be imposed within the proposed application and pipeline. Studies can involve other forms of data content, including merchandiser (*diser*) inventory lists, delivery schedules, among others. This holds promise for the development of future innovations aiding the entrepreneurs in pursuing more data-driven decisions and ensuring the competitive advantage and longevity with their products and services.

Acknowledgments. We would like to thank Ateneo de Manila University through the John Gokongwei School of Management (JGSOM) and the Business Insights Laboratory for Development (BUILD) for enabling our participation and future presentation of this study.

References

1. AnalyzeDocument - Amazon Textract, https://docs.aws.amazon.com/textract/

latest/dg/API_AnalyzeDocument.html

2. Claude 3 Haiku: our fastest model yet \ Anthropic, https://www.anthropic.com/news/claude-3-haiku
3. OCR Software, Data Extraction Tool - Amazon Textract - AWS, https://aws.amazon.com/textract/
4. Baier, M., Carballo, J.E., Chang, A.J., Lu, Y., Mojsilovic, A., Richard, M.J., Singh, M., Squillante, M.S., Varshney, K.R.: Sales-force performance analytics and optimization. IBM Journal of Research and Development **56**(6), 8:1–8:10 (Nov 2012). https://doi.org/10.1147/JRD.2012.2216314, https://ieeexplore.ieee.org/abstract/document/6353942, 7 citations (Crossref) [2024-10-13] Conference Name: IBM Journal of Research and Development
5. Buckland, M., Gey, F.: The relationship between Recall and Precision. Journal of the American Society for Information Science **45**(1), 12–19 (1994). https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L, https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199401%2945%3A1%3C12%3A%3AAID-ASI2%3E3.0.CO%3B2-L, 574 citations (Crossref) [2024-10-13] _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199401%2945%3A1%3C12%3A%3AAID-ASI2%3E3.0.CO%3B2-L
6. Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K.: Optical Character Recognition Systems. In: Chaudhuri, A., Mandaviya, K., Badelia, P., K Ghosh, S. (eds.) Optical Character Recognition Systems for Different Languages with Soft Computing, pp. 9–41. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-50252-6_2, https://doi.org/10.1007/978-3-319-50252-6_2
7. Chen, B., Zhang, Z., Langrené, N., Zhu, S.: Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review (Sep 2024). https://doi.org/10.48550/arXiv.2310.14735, http://arxiv.org/abs/2310.14735, arXiv:2310.14735
8. Church, K.W., Chen, Z., Ma, Y.: Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering **27**(6), 763–778 (Nov 2021). https://doi.org/10.1017/S1351324921000322, https://www.cambridge.org/core/journals/natural-language-engineering/article/emerging-trends-a-gentle-introduction-to-finetuning/C31D429D0928351D6A6692F8ECD1E7ED, 31 citations (Crossref) [2025-01-29]
9. Concepcion, R.S., Bedruz, R.A.R., Culaba, A.B., Dadios, E.P., Pascua, A.R.A.R.: The Technology Adoption and Governance of Artificial Intelligence in the Philippines. In: 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). pp. 1–10 (Nov 2019). https://doi.org/10.1109/HNICEM48295.2019.9072725, https://ieeexplore.ieee.org/abstract/document/9072725?casa_token=2xcX699p_fwAAAAA:UcwEKBOzHMKlXeNQg94Hj_R2SRV5Mwx3AxVbx7f39jHVklR6MDPLJUjUsxyGMIYWQfCV5S5Pz-0P, 21 citations (Crossref) [2024-10-18]
10. Devarasetty, N.: AI and Data Engineering: Harnessing the Power of Machine Learning in Data-Driven Enterprises. International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence **14**(1), 195–226 (Sep 2023), http://ijmlrcai.com/index.php/Journal/article/view/68, number: 1

11. Golin, M., Rauh, C.: The Impact of Fear of Automation (Dec 2022). https://doi.org/10.17863/CAM.91990, https://www.repository.cam.ac.uk/handle/1810/344569, publisher: Faculty of Economics, University of Cambridge
12. Gray, C.: Entrepreneurship, resistance to change and growth in small firms. Journal of Small Business and Enterprise Development **9**(1), 61–72 (Jan 2002). https://doi.org/10.1108/14626000210419491, https://doi.org/10.1108/14626000210419491, 139 citations (Crossref) [2025-01-18] Publisher: MCB UP Ltd
13. Habaradas, R., Jimenez, S., Mia, I.B., Javier, C., Enriquez, J., Ranieses, J.J.: Digital Technology Adoption Among Philippine Micro-, Small-,and Medium-Sized Enterprises: Barriers, Enablers & Challenges During COVID-19 **4** (Apr 2024). https://doi.org/10.61781/4-1I2024/1bmlm, 0 citations (Crossref) [2024-10-18]
14. Helo, P., Hao, Y.: Artificial intelligence in operations management and supply chain management: an exploratory case study. Production Planning & Control **33**(16), 1573–1590 (Dec 2022). https://doi.org/10.1080/09537287.2021.1882690, https://doi.org/10.1080/09537287.2021.1882690, 164 citations (Crossref) [2025-01-19] Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09537287.2021.1882690
15. Hirche, M., Greenacre, L., Nenycz-Thiel, M., Loose, S., Lockshin, L.: SKU performance and distribution: A large-scale analysis of the role of product characteristics with store scanner data. Journal of Retailing and Consumer Services **61**, 102533 (Jul 2021). https://doi.org/10.1016/j.jretconser.2021.102533, https://www.sciencedirect.com/science/article/pii/S0969698921000990, 7 citations (Crossref) [2024-10-13]
16. Huang, Y., Gursoy, D.: How does AI technology integration affect employees' proactive service behaviors? A transactional theory of stress perspective. Journal of Retailing and Consumer Services **77**, 103700 (Mar 2024). https://doi.org/10.1016/j.jretconser.2023.103700, https://www.sciencedirect.com/science/article/pii/S0969698923004514, 15 citations (Crossref) [2025-01-19]
17. Ilagan, J.B., Alabastro, Z.M., Basallo, C.L., Ilagan, J.R.: Exploratory Customer Discovery Through Simulation Using ChatGPT and Prompt Engineering. In: Yang, X.S., Sherratt, S., Dey, N., Joshi, A. (eds.) Proceedings of Ninth International Congress on Information and Communication Technology. pp. 51–60. Springer Nature, Singapore (2024). https://doi.org/10.1007/978-981-97-4581-4_5, 0 citations (Crossref) [2025-01-19]
18. Lamarine, M., Hager, J., Saris, W.H.M., Astrup, A., Valsesia, A.: Fast and Accurate Approaches for Large-Scale, Automated Mapping of Food Diaries on Food Composition Tables. Frontiers in Nutrition **5**, 38 (May 2018). https://doi.org/10.3389/fnut.2018.00038, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5954085/, 9 citations (Crossref) [2025-01-29]
19. Lee, D., Kim, S.C., Youn, I.H.: Development of a Ship's Logbook Data Extraction Model Using OCR Program. Journal of the Korean Society of Marine Environment & Safety **30**(1), 97–107 (2024). https://doi.org/10.7837/kosomes.2024.30.1.097, https://koreascience.kr/article/JAKO202410643409202.page, 0 citations (Crossref) [2024-10-13] Publisher: The Korean Society of Marine Environment and safety
20. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of Post-OCR Processing Approaches. ACM Comput. Surv. **54**(6), 124:1–124:37 (Jul 2021). https://doi.org/10.1145/3453476, https://dl.acm.org/doi/10.1145/3453476, 100 citations (Crossref) [2024-10-13]
21. Nguyen, T.H., Newby, M., Macaulay, M.J.: Information Technology Adoption in Small Business: Confirmation of a Proposed Framework. Journal of Small Business Management **53**(1), 207–227 (2015). https://doi.org/10.1111/jsbm.12058, https:

//onlinelibrary.wiley.com/doi/abs/10.1111/jsbm.12058, 151 citations (Crossref) [2025-01-19] _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsbm.12058
22. Perez, E., Kiela, D., Cho, K.: True Few-Shot Learning with Language Models. In: Advances in Neural Information Processing Systems. vol. 34, pp. 11054–11070. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/hash/5c04925674920eb58467fb52ce4ef728-Abstract.html
23. Premkumar, G., Roberts, M.: Adoption of new information technologies in rural small businesses. Omega **27**(4), 467–484 (Aug 1999). https://doi.org/10.1016/S0305-0483(98)00071-1, https://www.sciencedirect.com/science/article/pii/S0305048398000711, 570 citations (Crossref) [2024-10-13]
24. Raj, B.S., Sujith, B., Krishna, J.V., Anjali: Seamless Attendance Logging Application: A Technological Revolution in Logbook Capture. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). pp. 1–6 (Apr 2024). https://doi.org/10.1109/I2CT61223.2024.10543455, https://ieeexplore.ieee.org/abstract/document/10543455, 0 citations (Crossref) [2024-10-13]
25. Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing Generative Models via Precision and Recall. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html
26. Simon, R., Suarez, M.T.: Examining the Behavioral Intention of Philippine MSMEs Toward Business Intelligence Adoption. Journal of Business and Management **28**(1), 67–99 (Jan 2022). https://doi.org/10.1504/JBM.2022.141295, https://www.inderscienceonline.com/doi/abs/10.1504/JBM.2022.141295, 0 citations (Crossref) [2024-10-18] Publisher: Inderscience Publishers
27. Szymański, P., Kajdanowicz, T.: A scikit-based Python environment for performing multi-label classification (Dec 2018). https://doi.org/10.48550/arXiv.1702.01460, http://arxiv.org/abs/1702.01460, arXiv:1702.01460 [cs]
28. Tolbert, F.: Why Point of Sales Data Matters for Demand Management. The Journal of Business Forecasting **27**(4), 33–35 (2008), https://www.proquest.com/docview/226917513/abstract/94C35C4D88CE47EFPQ/1, num Pages: 3 Place: Flushing, United States Publisher: Journal of Business Forecasting
29. Windrum, P., de Berranger, P.: The adoption of e-business technology by SMEs