DOMAIN ADAPTATION REGULARIZED LAYOUTLM WITH AUTOMATIC DOMAIN DISCOVERY FROM TOPIC MODELLING

Chen Lin and Piush Kumar Singh and Yourong Xu and Eitan Lees
and Rachna Saxena and Sasidhar Donaparthi and Hui Su

Fidelity Investments, 245 Summer Street, Boston, MA 02210

ABSTRACT

In this paper, we propose using domain adaptation to improve the generalizability and performance of LayoutLM, a pre-trained language model that incorporates layout information of a document image. Our approach uses topic modelling to automatically discover the underlying domains in a document image dataset where domain information is unknown. We evaluate our approach on the challenging RVL-CDIP dataset and demonstrate that it significantly improves the performance of LayoutLM on this dataset. Our approach can be applied to other NLP models to improve their generalization capabilities, making them more applicable in real-world scenarios, where data is often collected from a variety of domains.

KEYWORDS

LayoutLM, Domain Adaptation, Automatic Domain Discovery, Topic Modelling, RVL-CDIP

1. INTRODUCTION

In recent years, natural language processing (NLP) has seen remarkable progress, with many pre-trained language models achieving state-of-the-art performance in various tasks [1], [2], [3]. However, these models often struggle to generalize well to new domains, especially when the target domain data is out-of-distribution, which can limit their applicability in real-world scenarios. Domain adaptation [4], [5], [6] has emerged as a promising approach to address the issue of limited generalizability of models to new domains. This technique involves regularizing the models to focus on domain-independent features, thereby improving their ability to perform well on unseen data. This is important because in many real-world applications, it is often not possible to collect a large amount of labelled data for the target domain, or the data distribution in the target domain may shift over time. By using domain adaptation, it is possible to transfer knowledge from the source domain, which has a larger and more diverse dataset, to the target domain, thereby improving the performance of the model on the target domain.

Recent research on domain adaptation methods has focused on neural network-based models that learn domain-invariant feature representations. One such model is the updated version of structural correspondence learning (SCL) [7], [8], which creates new features based on linear classifiers' weights that predict feature values given other features. In neural SCL, a multi-layer perceptron is used to predict feature values. While successful [9], [10], these approaches still rely on traditional feature engineering and require the practitioner to decide which parts of the feature space are the predictors and which should be predicted. Another line of approaches is to work

David C. Wyld et al. (Eds): AIBD, MLSC, ACSTY, NATP, CCCIoT, SVC, SOFE, ITCSS -2024
pp. 279-286, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140422

directly on word inputs passed through an embedding layer [11], [12], the most successful one is the domain adversarial neural network (DANN) [12]. DANN trains a neural network that can predict the label of interest but cannot distinguish between the two domains from the feature set. The goal is to learn domain-independent cues while discarding domain-specific noise.

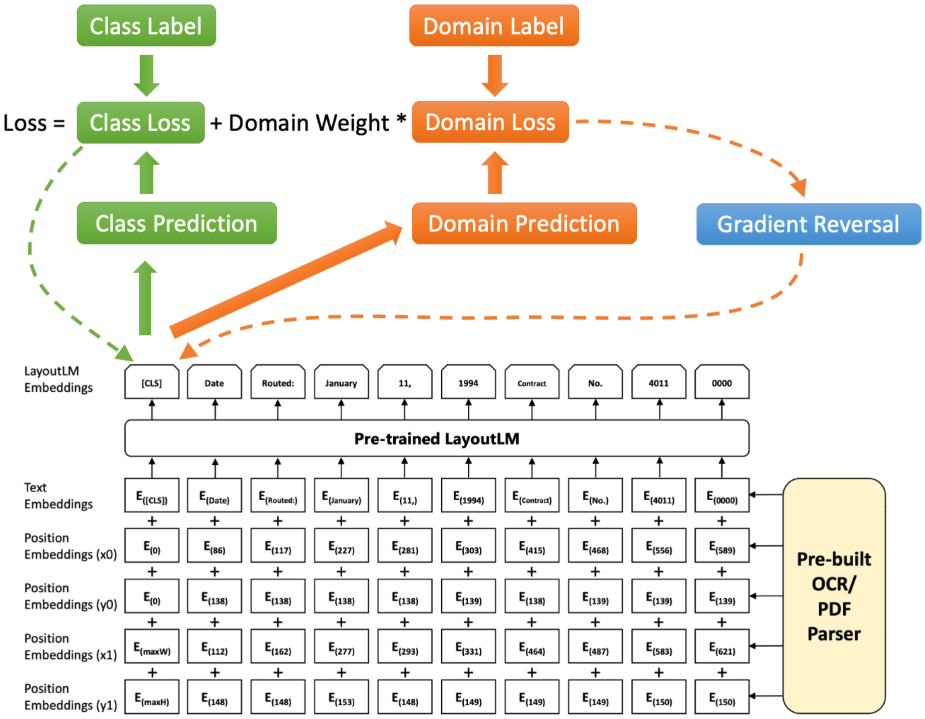
Class Label
Domain Label
Loss = Class Loss + Domain Weight * Domain Loss
Class Prediction
Domain Prediction
Gradient Reversal
LayoutLM Embeddings
[CLS]
Date
Routed:
January
11,
1994
Contract
No.
4011
0000
Pre-trained LayoutLM
Text Embeddings
$E_{([CLS])}$ $E_{(Date)}$ $E_{(Routed:)}$ $E_{(January)}$ $E_{(11,)}$ $E_{(1994)}$ $E_{(Contract)}$ $E_{(No.)}$ $E_{(4011)}$ $E_{(0000)}$
Position Embeddings (x0)
$E_{(0)}$ $E_{(86)}$ $E_{(117)}$ $E_{(227)}$ $E_{(281)}$ $E_{(303)}$ $E_{(415)}$ $E_{(468)}$ $E_{(556)}$ $E_{(589)}$
Position Embeddings (y0)
$E_{(0)}$ $E_{(138)}$ $E_{(138)}$ $E_{(138)}$ $E_{(138)}$ $E_{(139)}$ $E_{(138)}$ $E_{(139)}$ $E_{(139)}$ $E_{(139)}$
Position Embeddings (x1)
$E_{(maxW)}$ $E_{(112)}$ $E_{(162)}$ $E_{(277)}$ $E_{(293)}$ $E_{(331)}$ $E_{(464)}$ $E_{(487)}$ $E_{(583)}$ $E_{(621)}$
Position Embeddings (y1)
$E_{(maxH)}$ $E_{(148)}$ $E_{(148)}$ $E_{(153)}$ $E_{(148)}$ $E_{(149)}$ $E_{(149)}$ $E_{(149)}$ $E_{(150)}$ $E_{(150)}$
Pre-built OCR/ PDF Parser

Figure 1. Model structure of applying domain adaptation on LayoutLM.

One promising model for NLP tasks is pre-training of text and layout for document image understanding (LayoutLM) [13], [14], [15], [16]. It is a pre-trained language model that incorporates both textual and layout information. To pre-train the model, a large dataset of text-layout pairs is used, where the text represents the content of a document, and the layout represents the visual arrangement of the text and other elements on the page. This pre-training allows the model to learn the relationship between text and layout, and to understand the document structure and element arrangement. The model can then be fine-tuned for specific supervised tasks but may still suffer from domain-specific biases and struggle to generalize. To address this, we propose using DANN as a regularizer in the fine-tuning stage of LayoutLM to improve its cross-domain performance. While DANN-enhanced fine-tuning has been tested for other pre-trained language models [17], [18], [19], to our knowledge, it has not been studied for the LayoutLM model. In this paper, we propose incorporating DANN in the fine-tuning stage of LayoutLM to improve its generalizability and performance.

We implement an automatic domain discovery approach based on topic modelling [20] or other clustering methods [21] to learn the domain information for an unseen dataset for DANN training. Topic modelling is a method in natural language processing that involves automatically identifying the topics present in a corpus of text documents. It is a form of unsupervised machine learning, where the model is not provided with any labelled data but must instead learn to identify the topics present in the text on its own. This is done by analysing the words and phrases used in the documents and grouping them into clusters based on their co-occurrence patterns. The goal of topic modelling is to uncover the hidden structure in a collection of documents, and to identify the latent topics that are present in the text. This approach aims to identify the underlying domains present in a dataset.

To evaluate the effectiveness of our proposed approach, we conduct experiments on the RVL-CDIP dataset [22], a public dataset of document images with 16 classes. The dataset is commonly used for document classification tasks and contains a diverse range of document types, making it a challenging dataset for domain adaptation. Through our experiments, we demonstrate that our proposed approach can significantly improve the performance of LayoutLM on the RVL-CDIP dataset. In summary, this paper introduces a novel approach that leverages automatic domain discovery and DANN in fine-tuning LayoutLM. Our approach is evaluated on the challenging RVL-CDIP dataset, and the results demonstrate significant performance gains.

2. METHOD

2.1. OCR

To pre-process the document images in the RVL-CDIP dataset, we used Optical Character Recognition (OCR) to extract the text content from the images. Specifically, we used Tesseract, an open-source OCR engine to perform OCR on the images. Tesseract is a widely used OCR engine that supports over 100 languages and has been shown to achieve state-of-the-art performance on several benchmarks [23].

2.2. Topic Modelling

To represent different domains, we performed Topic Modelling [20] of the OCRed text from the RVL-CDIP images with BERT embeddings [2]. We adopted BERT embeddings specifically to capture the nuanced semantic attributes inherent in the text. These embeddings served as the foundation for our clustering, implemented via two prominent clustering techniques, K-means [24] and Latent Dirichlet Allocation (LDA) [25]. Our choice of these methods was rooted in their capacity to capture the essence of text data effectively. Further to ascertain the optimal number of clusters, which is pivotal for the efficacy of our approach, we turned to the DUNN index [26]. The DUNN index, for clarity, measures the ratio of the minimal inter-cluster distance to the maximal intra-cluster distance. It's a trusted metric in the clustering domain that provides insights into the quality of clustering by ensuring that clusters are compact and well-separated. Our deployment of this metric testifies to the methodological rigor we have applied to ensure that our model is both accurate and justifiable.

2.3. DANN for Fine-Tuning Layoutlm

Figure 1 shows the DANN regularized LayoutLM network, which takes parsed tokens and their position embeddings from OCR as input. We use LayoutLM v1 [13] for this work because it is well-performing and has a permissive license. For each instance, both the class label and the domain information are provided, and the network encodes the input to predict both the task label (one of 16 image classes) and the domain that the instance belongs to. The loss function combines the losses from both the task label and the domain prediction. To prevent the network from distinguishing between domains, a gradient reversal layer is added between the learned representation layer and the domain prediction layer. During the forward pass, the gradient reversal layer simply passes its input forward, while during the backward pass, the gradients are multiplied by -1, making it more challenging to distinguish between domains. This process enhances domain-independent feature representations while filtering out domain-specific noises.

2.4. Training and Testing

We fine-tuned the LayoutLM with DANN regularization on the pre-set RVL-CDIP training set, validated it on the validation set, and tested it on the test set. The domain and class information were not given to the model during testing. We evaluated the model's performance using standard metrics such as precision, recall, and F1-score.

2.5. Experimental Setup

We used the Hugging Face implementation of LayoutLM for our experiments. We fine-tuned the model on the RVL-CDIP dataset using the modified loss function and domain adaptation method. We used the Adam optimizer with a learning rate of 2e-5 and a batch size of 32. We trained the model for 6 epochs with early stopping and selected the best model based on the validation set's performance.

2.6. Evaluation

We evaluated the performance of our proposed approach using the standard evaluation metrics mentioned above. We compared our results with the baseline LayoutLM model without Domain Adaptation and with randomly assigned clusters.

3. RESULTS

Our experimental results show that domain adaptation significantly improves the performance of LayoutLM on the RVL-CDIP dataset. The results of our experiments are shown in Table 1. Without domain adaptation, LayoutLM achieved an F1 score of 81.3%. Applying our proposed approach on randomly generated clusters, the F1 score is similar, 81.4%. After applying our proposed approach on carefully selected clusters of two, four, or eight, the accuracy improved to 82.4%, 81.7%, and 82.1% respectively, representing the biggest relative improvement of 1.0% (between our method on 2 clusters and on random clusters). This improvement is statistically significant with a p-value of 0.00214, as determined by Wilcoxon signed-rank test [27]. The Precision, Recall, and F1 in Table 1 are macro average results of the 16 image types in the test set of RVL-CDIP.

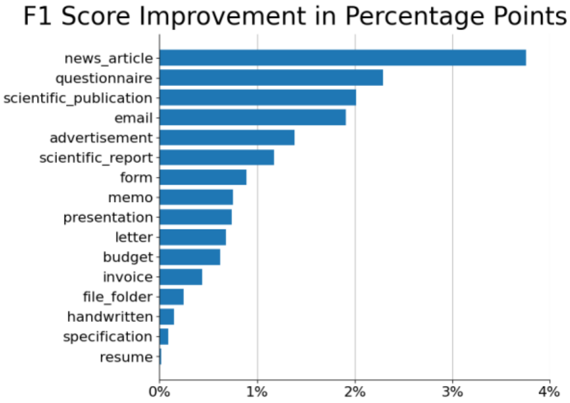

Figure 2. F1 improvement in percentage points across 16 classes for DA on 2 clusters vs. DA on random clusters.

Figure 2 offers a comparative analysis of the F1 score enhancement, expressed in percentage points, across all 16 categories. The comparison is made between the top-performing domain adaptation applied to two selected clusters and random clusters. Based on Wilcoxon signed-rank test, a performance gain of 1% is considered significant. From this perspective, several categories exhibited substantial improvements. The ’Scientific Report’ category demonstrated a gain of 1.17%, ’Advertisement’ improved by 1.39%, ’Email’ saw a growth of 1.91%, ’Scientific Publication’ rose by 2.02%, ’Questionnaire’ increased by 2.29%, and ’News Article’ stood out with a remarkable enhancement of 3.76%.

4. Discussion

Our proposed approach for incorporating DANN in the fine-tuning stage of LayoutLM and using automatic domain discovery based on topic modeling or other clustering methods has demonstrated significant improvements in the generalizability and performance of LayoutLM on the RVL-CDIP dataset.

However, one interesting aspect that needs to be discussed is the effect of the number of clusters on the performance of the proposed method. In our experiments, we found that using a large number of clusters (i.e., more than 50) did not significantly improve the performance, while using a smaller number of clusters (i.e., 2-8) produced the best results. The decrease in performance could be because a larger number of clusters may be too specific and thus introduce noise and make it more challenging for the model to generalize to new domains.

Our base LayoutLM results on the RVL-CDIP test set fell short of the best-reported results, potentially due to differing fine-tuning setups or hardware specifics. Limited by a single regular GPU for cost efficiency, we faced constraints on batch size and memory. Despite needing further investigation to understand this performance gap, our focus wasn’t achieving state-of-the-art results, but rather to showcase our method’s effectiveness in enhancing LayoutLM performance, validated by a statistically significant p-value.

Our study demonstrates the effectiveness of domain adaptation when applied to clusters within the diverse RVL-CDIP dataset, which comprises 16 distinct classes. We propose that the blurred domain information across these classes refers not to the class details but to the complex visual features within the data.

We hypothesize that maximum benefit from a domain adaptation regularized LayoutLM would be derived from samples that exhibit significantly different structural or visual characteristics compared to the mainstream training data. These include:

- Advertisements and News Articles: These display a rich blend of text, graphics, and images in diverse layouts.
- Forms and Questionnaires: Characterized by distinct structural elements for input segregation.
- Presentations: These are multifaceted, incorporating text, diagrams, charts, and images.
- Scientific Reports/Publications: These complex documents house diagrams, charts, tables, and possibly images or illustrations.

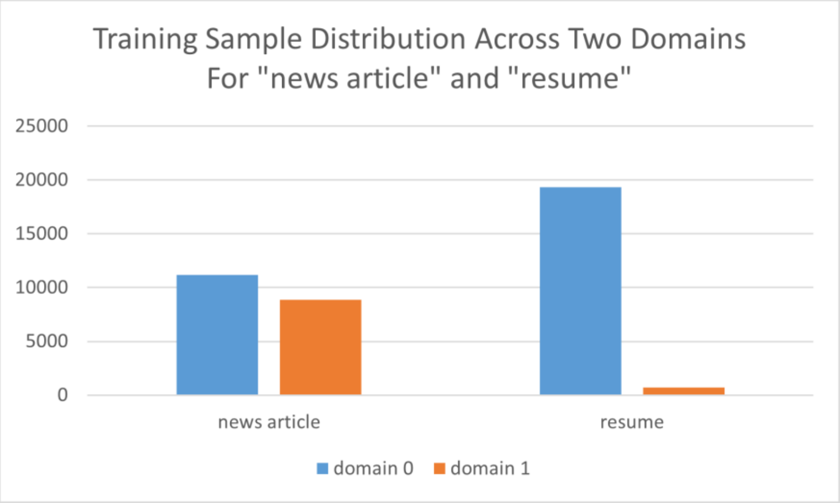

Figure 3. Training Sample Distribution Across Two Domains for "news article" and "resume".

The F1 score improvements across the 16 classes substantiate our hypothesis. Most improved were "news articles", "questionnaire", "scientific publication/reports", and "advertisement" - all typified by complex visual features. Emails, too, showed significant improvement, possibly due to irregular layouts, unique fonts, and distinct signatures. Despite not exceeding the 1% improvement threshold, "form", "memo" and "presentation" classes showed substantial F1 increments. This outcome indicates the effectiveness of our domain adaptation technique in managing complex visual features in diverse document types.

Table 1. Performance of LayoutLM with Domain Adaptation (DA) on different numbers of clusters vs. Base and DA on randomly generated clusters.

Model	**Precision**	**Recall**	**F1**
Base LayoutLM	0.820	0.812	0.813
DA on random clusters	0.824	0.813	0.814
DA on 2 clusters	**0.832**	**0.823**	**0.824**
DA on 4 clusters	0.822	0.815	0.817
DA on 8 clusters	0.828	0.821	0.821

For classes like "handwritten", "specification", and "resume" that saw less improvement, the uniformity in their topics and visual complexity might have made domain adaptation less advantageous. Figure 3 illustrates the distribution of training samples across two identified domains for the most improved class, "news article", and the least improved, "resume". The "resume" class exhibits a more homogeneous distribution with most of its samples (96.38%) clustered in one domain. Conversely, "news article" samples are nearly evenly distributed across both domains (11,160 vs. 8,851), indicating a broader diversity in topics or themes.

Our proposed approach can also be used in other applications that involve domain adaptation for text or image classification tasks. By using automatic domain discovery to identify the underlying domains in a dataset, our method can help improve the performance of models on new domains, even when the domain information is unknown.

5. Conclusion

In summary, we used Topic Modeling and Domain Adaptation to improve the generalizability and performance of LayoutLM on the challenging RVL-CDIP dataset. We modified the loss function to include domain loss and used reverse gradient propagation to regularize the model's domain-specific features. Our approach achieved significantly improved results on the test set, demonstrating its effectiveness in improving the generalization capabilities of LayoutLM models.

Acknowledgements

The authors are solely accountable for the content, and it does not reflect the official stance of Fidelity Investments. We are grateful to the reviewers for their insightful feedback and constructive criticism. We also appreciate the support of Victor Lo, Melissa Nysewander, and Alec Bethune for this work.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[4] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo *et al.*, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[5] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.

[7] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 120–128.

[8] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 440–447.

[9] R. Levy and L. Specia, "Proceedings of the 21st conference on computational natural language learning (conll 2017)," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017.

[10] T. Miller, "Simplified neural unsupervised domain adaptation," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2019. NIH Public Access, 2019, p. 414.

[11] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[13] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1192–1200.

[14] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che *et al.*, “Layoutlmv2: Multi-modal pre-training for visually-rich document understanding,” *arXiv preprint arXiv:2012.14740,* 2020.

[15] Q. Xu, L. Wang, H. Liu, and N. Liu, “Layoutlm-critic: Multimodal language model for text error correction of optical character recognition,” in *Artificial Intelligence and Robotics: 7th International Symposium, ISAIR 2022, Shanghai, China, October 21-23, 2022, Proceedings, Part II*. Springer, 2022, pp. 136–146.

[16] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, “Layoutlmv3: Pre-training for document ai with unified text and image masking,” in *Proceedings of the 30th ACM International Conference on Multimedia,* 2022, pp. 4083–4091.

[17] X. Han and J. Eisenstein, “Unsupervised domain adaptation of contextualized embeddings for sequence labeling,” *arXiv preprint arXiv:1904.02817,* 2019.

[18] C. Lin, S. Bethard, D. Dligach, F. Sadeque, G. Savova, and T. A. Miller, “Does bert need domain adaptation for clinical negation detection?” *Journal of the American Medical Informatics Association,* vol. 27, no. 4, pp. 584–591, 2020.

[19] S. Gururangan, A. Marasovic´, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” *arXiv preprint arXiv:2004.10964,* 2020.

[20] I. Vayansky and S. A. Kumar, “A review of topic modeling methods,” *Information Systems,* vol. 94, p. 101582, 2020.

[21] T. S. Madhulatha, “An overview on clustering methods,” *arXiv preprint arXiv:1205.1117,* 2012.

[22] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval,” in *International Conference on Document Analysis and Recognition (ICDAR),* 2015.

[23] R. Smith, “An overview of the tesseract ocr engine,” in *Ninth international conference on document analysis and recognition (ICDAR 2007),* vol. 2. IEEE, 2007, pp. 629–633.

[24] A. Ahmad and L. Dey, “A k-mean clustering algorithm for mixed numeric and categorical data,” *Data & Knowledge Engineering,* vol. 63, no. 2, pp. 503–527, 2007.

[25] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey,” *Multimedia Tools and Applications,* vol. 78, pp. 15 169–15 211, 2019.

[26] J. C. Bezdek and N. R. Pal, “Cluster validation with generalized dunn’s indices,” in *Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems*. IEEE, 1995, pp. 190–193.

[27] R. F. Woolson, “Wilcoxon signed-rank test,” *Wiley encyclopedia of clinical trials,* pp. 1–3, 2007.

AUTHORS

Chen Lin is a Principal Data Scientist at Fidelity Investments, developing AI solutions for intelligent document understanding.

Piush Kumar Singh is a Data Scientist at Fidelity Investments.

Yourong Xu is a Principal Data Scientist at Fidelity Investments.

Eitan Lees is a Data Engineer at Fidelity Investments.

Rachna Saxena is a Data Scientist at Fidelity Investments.

Sasidhar Donaparthi is a Data Scientist at Fidelity Investments.

Hui Su is a Vice President, Data Science Practice Lead at Fidelity Investments.