Applications in Explainable Text Identification and Classification in Multilingual Document Processing

Shalini Puri[1], Midhunchakkaravarthy Janarthanan[2], Ganesh Khekare[3]
[1]Lincoln University College Malaysia; [2]Lincoln University College Malaysia; [3]Vellore Institute of Technology, Vellore, India
eng.shalinipuri30@gmail.com
midhun@lincoln.edu.my
khekare.123@gmail.com

Abstract: Multilingual document classification has grown in importance as information systems have become more global. Despite their effectiveness, classic AI models frequently operate as opaque black boxes; efforts are being made to increase the models' explainability. Fuzzy logic and membership functions are used to model linguistic ambiguity and uncertainty, enabling interpretable feature representation and reasoning. This paper presents several applications of multilingual text extraction and classification using XAI and machine learning. It provides a contextual foundation for XAI-based text identification in multilingual document classification in diverse application areas. Furthermore, it addresses various applications, highlighting the languages used and the document type.

Keywords: Explainability; Document Processing; Sustainable Learning; Multilingualism; Applications.

Introduction

Due to numerous social, economic, and cultural factors, multilingual picture document categorization has emerged as a research area in the current era. By helping international corporations streamline processes, including marketing and documentation, across linguistically disparate regions, it promotes economic globalization. By serving a worldwide audience with a variety of linguistic preferences, it expands market reach [9]. Two of the most common tasks in natural language processing are text categorization and sentiment analysis, which have several new applications in diverse fields, including multilingual document processing, healthcare, policy making, etc. Text categorization is a key job in Natural Language Processing (NLP) that enables automatic document classification for applications such as sentiment analysis, intent detection, customer feedback analysis, and educational evaluation [10].

In contrast to plain-text document classification, document image classification involves categorizing documents based on their structure and content, including emails, forms, and other similar documents [6]. These days, effective bilingual systems for categorizing both non-image and picture materials are required due to the fast digitization. Although many bilingual NLP and image-based systems provide solutions for real-world problems, they primarily focus on text extraction, identification, and recognition

tasks with limited document types [4]. These days, the use of Explainable Artificial Intelligence (XAI) in text processing has significantly improved the systems' performance, interpretability, and reliability. Effective explainable classification becomes more complex due to the enormous amounts of multilingual, domain-diverse textual data produced by the exponential rise of social media and internet-based content [14].

While Deep Learning (DL) and Conventional Artificial Neural Network (ANN) models often achieve high accuracy, they function as "black boxes," offering no insight into their decision-making processes. However, Fuzzy Set Theory, which was developed to address ambiguity and uncertainty, has shown itself to be very useful in handling incomplete and imprecise data in several domains [2] [3] [6]. Although Deep Neural Networks (DNNs) have achieved remarkable performance improvements in document and image classification, little research has been done in this area that explores the explainability of these models [13]. Similarity metrics and rule extraction techniques, which improve interpretability, are crucial to fuzzy-based text classification; however, existing hybrid models quickly suffer from slow convergence, sensitivity to noise, and decreased effectiveness on incomplete multilingual datasets [8].

This study presents a discussion on several such applications from state-of-the-art and also analyzes them in comparison. There are significant applications concerning text extraction and processing, such as the processing of news articles, spam emails, semantic analysis, multilingual images, and legal document verification, for example, practical industrial scenarios where warnings and alerts are displayed in multiple languages. The paper structure is as follows: The next section presents a contextual foundation for XAI-based text identification in multilingual document classification in diverse application areas. The subsequent section tabulates various applications year by year, highlighting the languages used and the document type. Finally, the paper concludes with a discussion of future extensions.

Contextual Foundation of Explainable Text Identification

The review [1] proposed XAI-based predictive coding and simple explainable predictive coding methods to locate responsive snippets for actual legal document reviews. This work [1] was expanded in [2] to aid snippet search response within responsive documents and to enhance the effectiveness and quality of legal document assessment. Another work [3] presented an XAI method for image captioning using DL, providing a visual link between the region of objects in the images and the specific words (or phrases) in the generated sentences. It was evaluated using MSCOCO and Flickr30K datasets. Another article [4] discussed advanced applications of bilingual document analysis systems based on feature extraction techniques, document sets, classifiers, and accuracy for English-Hindi and other language pairs. The next article presented a bi-level image classifier to categorize printed and handwritten English documents by following preprocessing, segmentation, feature extraction, SVM-based optical character recognition, word association, and fuzzy matching-based document classification.

Another contribution [6] proposed two multilingual datasets, WIKI-DOC and MULTIEURLEX-DOC, for document image classification, like multi-label classification, and zero-shot cross-lingual transfer. Another work [7] presented an interpretable Bengali multiclass news classification with 9 classes, achieving an

accuracy of 92%. On the other hand, the interpretable text classifier [8] identified legal document review in construction delay dispute matters to enhance document classification accuracy by using delay-related snippets and their perspectives. The cost-efficient DNN-based multilingual image document classification [9] employed five new activation functions, while the multilingual text categorizer and sentiment analyzer utilized Bidirectional Encoder Representations from Transformers (BERT) and zero-shot techniques to classify Twitter data. The next contribution [11] selected features using SHapley Additive exPlanations (SHAP) to classify SMS spam in Dravidian languages, achieving real-time accuracy of 90-92%.

Another work [12] classified the artistic images using XAI and fuzzy Techniques. When compared to multitask learning, its context-aware features produced results that were up to 19% more accurate utilizing the residual network architecture and 3% more accurate using ConvNeXt. Another work [13] classified document images using ten DL Models and two datasets, RVL-CDIP and Tobacco3482. It was observed that the Occlusion and DeepSHAP techniques provided the best explanations using the DeepSHAP. A review [14] discussed multilingual document classification using XAI, fuzzy logic, and Convolutional Neural Networks (CNN), and presented a set of research questions. Another work [15] classified products and services published in the descriptions of public procurement tenders, such as business documents, using BERT. It was pre-trained on an English dataset and evaluated on the Czech dataset extracted from the Gazette of Public Procurement of the Czech Ministry of Local Development. On the other hand, the explainable multilingual text document classification [16] model applied fuzzy logic and XAI while maintaining interpretability using an iterative optimization and evaluation loop.

Applications and Practical Significance

The proposed work highlights the applications of XAI and machine learning techniques, such as fuzzy logic and CNN, towards multilingual document classification. It has primarily five categories of document application areas, as shown in Table 1, including legal documents [1] [2] [8] [10], business documents [15], images [3] [5] [6] [9] [13] and painting images [12], news articles [7] [16], and spam data [11]. These works were performed in various languages, including English, European, Bengali, Japanese, German, French, Chinese, Spanish, Hindi, Telugu, Kannada, and many more. The application categories included XAI-based model versions for text classification, image captioning and classification, news classification, and SMS spam detection. Some other applications did not apply XAI for classification, such as printed and handwritten document classification, as well as multilingual text categorization and sentiment analysis.

Table 1. Application areas, languages, and document types in language-based text classification.

Ref. No.	Year	Application Area	Language	Document Type
[1]	2018	XAI-based Text Classification	English	Legal Documents
[2]	2019	XAI-based Text Classification	English	Legal Documents
[3]	2019	XAI-based image captioning	-	Images

[5]	2020	Fuzzy Matching-based Classifier for Printed and Handwritten Documents	English	Images
[6]	2023	Document Image Classification	European	Images
[7]	2023	XAI-based news classification	Bengali	News Articles
[8]	2023	XAI-based text classification for construction delay disputes	English	Legal Documents
[9]	2023	DNN for document classification	Multilingual	Images
[10]	2023	Multilingual text categorization and sentiment analysis	English, Japanese, German, French, Chinese, and Spanish	Twitter Data
[11]	2024	XAI-based SMS spam classification for Dravidian languages	English, Hindi, Telugu & Kannada	Spam Data
[12]	2024	XAI-based DL for artistic images using fuzzy	-	Paintings-Images
[13]	2024	Document image classification using DL	English	Images
[15]	2025	Multilingual classification using DL	Multilingual (104 Languages)	Business Documents
[16]	2026	Text document classification using fuzzy logic and XAI	Multilingual	News Articles

Conclusions and Prospective Horizons

The presented paper investigated the applications of automatic classification of multilingual document text using XAI and DL techniques. Several applications of XAI and machine learning for multilingual text extraction and categorization were shown in this work. In a variety of application domains, it presented a contextual basis for XAI-based text identification in multilingual document classification. It also discussed different applications, emphasizing the document type and the languages utilized. This work can be further extended to include agentic AI for text extraction security, interpretability, and trustworthiness.

References

1. R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang and H. Zhao, "Explainable text classification in legal document review a case study of explainable predictive coding," IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 1905-1911, 2018. https://doi.org/10.1109/BigData.2018.8622073

2. C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall and H. Zhao, "A framework for explainable text classification in legal document review," IEEE International Conference on Big Data (Big Data),” Los Angeles, CA, USA, pp. 1858-1867, 2019. https://doi.org/10.1109/BigData47090.2019.9005659
3. H. Han, S. Kwon and J. Choi, “EXplainable AI (XAI) Approach to Image Captioning,” *The Journal of Engineering*, vol. 2020, issue 13, pp. 589-594, 2020. https://doi.org/10.1049/joe.2019.1217
4. S. Puri and S. P. Singh, “Advanced Applications on Bilingual Document Analysis and Processing Systems,” *International Journal of Applied Metaheuristic Computing*, vol. 11, issue 4, pp. 149-193, 2020. https://doi.org/10.4018/IJAMC.2020100108
5. S. Puri and S. P. Singh, “A Fuzzy Matching based Image Classification System for Printed and Handwritten Text Documents,” *Journal of Information Technology Research*, vol. 13, issue 2, pp. 155-194, 2020. https://doi.org/10.4018/JITR.2020040110
6. Y. Fujinuma, S. Varia, N. Sankaran, S. Appalaraju, B. Min and Y. Vyas, “A Multi-Modal Multilingual Benchmark for Document Image Classification,” in Findings of the *Association for Computational Linguistics: EMNLP 2023*, Singapore, Association for Computational Linguistics, pp. 14361–14376, 2023. https://aclanthology.org/2023.findings-emnlp.958/
7. M. F. Sikder et al., "Explainable Bengali multiclass news classification," 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, pp. 1-6, 2023. https://doi.org/10.1109/ICCIT60459.2023.10441218
8. N. Huber-Fliflet et al., "Explainable text classification for legal document review in construction delay disputes," IEEE International Conference on Big Data (BigData), Sorrento, Italy, pp. 1928-1933, 2023. https://doi.org/10.1109/BigData59044.2023.10386240
9. S. Banerjee and D. Shende, "MLing-Net: A computationally inexpensive deep neural framework designed to perform multilingual image document classification," 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, pp. 1-6, 2023. https://doi.org/10.1109/IEMENTech60402.2023.10423510
10. G. Manias, A. Mavrogiorgou, A. Kiourtis et al., “Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying Twitter data,” *Neural Comput & Applic*, vol. 35, pp. 21415–21431, 2023. https://doi.org/10.1007/s00521-023-08629-3
11. K. G. Thirumalai, K. S. Prakash, A. M. Abirami, E. Ramanujam and S. Sumitra, "XAI-based feature selection for sms spam classification in Dravidian languages," 5th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, pp. 1-6, 2024. https://doi.org/10.1109/ICITIIT61487.2024.10580065
12. J. Fumanal-Idocin, J. Andreu-Perez, O. Cordón, H. Hagras and H. Bustince, "ARTxAI: Explainable Artificial Intelligence Curates Deep Representation Learning for Artistic Images Using Fuzzy Techniques," IEEE Transactions on Fuzzy Systems, vol. 32, no. 4, pp. 1915-1926, 2024. https://doi.org/10.1109/TFUZZ.2023.3337878
13. Saifullah, S. Agne, A. Dengel and S. Ahmed, "The Reality of High Performing Deep Learning Models: A Case Study on Document Image Classification," IEEE Access, vol. 12, pp. 103537-103564, 2024. https://doi.org/10.1109/ACCESS.2024.3425910

14. S. Puri, M. Janarthanan and G. Khekare, “Multilingual document classification using XAI: a review,” International Conference LGPR, SGS - Engineering & Sciences, vol. 1, issue 2, pp. 1-4, 2025. https://spast.org/techrep/article/view/5397
15. P. Bednár, J. I. Vanko and E. Žiak, "Deep learning methods for multilingual classification of business documents," IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI), Stará Lesná, Slovakia, pp. 000327-000330, 2025. https://doi.org/10.1109/SAMI63904.2025.10883330
16. S. Puri, M. Janarthanan and G. Khekare, “A novel approach to multilingual text document classification using fuzzy logic and XAI,” International Conference LGPR, SGS - Initiative, vol. 1, issue 1, pp. 1-7, 2026.