RESEARCH-ARTICLE

From Unstructured Documents to Annotated Information: An Optimized Pipeline to Process Industrial Requirements

NOCENTE ARIANNA, University of Pisa, Pisa, PI, Italy

RISI RICCARDO, Columbia Business School, New York, NY, United States

PANNOCCHIA GABRIELE, University of Pisa, Pisa, PI, Italy

ROSSETTI GIULIO, Italian National Research Council, Rome, RM, Italy

Published: 13 December 2024

NLPIR 2024: 2024 8th International Conference on Natural Language Processing and Information Retrieval
December 13 - 15, 2024
Okayama, Japan

NLPIR '24: Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval (December 2024)
https://doi.org/10.1145/3711542.3711576
ISBN: 9798400717383

From Unstructured Documents to Annotated Information: An Optimized Pipeline to Process Industrial Requirements

Nocente Arianna
Department of Information Engineering
University of Pisa
Pisa, Italy
Alstom Ferroviaria Spa
Bologna, Italy
arianna.nocente@phd.unipi.it

Risi Riccardo
Marketing Division
Columbia Business School
New York, NY, USA
rr3566@columbia.edu

Pannocchia Gabriele
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy
gabriele.pannocchia@unipi.it

Rossetti Giulio
Institute of Science and Technologies
Italian National Research Council (CNR)
Pisa, Italy
giulio.rossetti@isti.cnr.it

Abstract

Managing bid documentation in large, evolving technology companies is inherently complex, often due to inconsistencies in information such as translations, file updates, and manual data extraction. These processes involve multiple departments, including software, hardware, products, infrastructure, materials, and regulations, requiring collaboration across geographically distributed teams with different native languages. This complexity is exacerbated by the need to trace requirements from bid offers to code and product development, and to perform similarity analysis when needed. Unstructured information comes from diverse sources like scans and/or editable texts with tables and images, written in various languages and using domain-specific terminology. Manual processing is error-prone, and translating data can lead to the loss of context-specific meanings or issues in safety-critical domains. This study combines Natural Language Processing (NLP) and Optical Character Recognition (OCR) to classify data into "information" or "requirement" while preserving multilingualism. A dual-pipeline approach is developed, featuring both a meta-classifier (an ensemble of Logistic Regression, Support Vector Machine, Multinomial Naive Bayes, and Random Forest) for robust and interpretable results, and a BERT model for capturing subtle linguistic patterns. The proposed pipeline is validated using a real-world case study in railway requirement annotation. Additionally, to demonstrate the methodology's flexibility, a second case study is conducted on topic classification of newspaper articles using publicly accessible data. The pipeline's output is a software solution that uses pre-trained models tailored to the respective domains. Future developments will include the creation of a graphical user interface (GUI), enabling distributed users to easily and efficiently search, update their requirements, and extract custom PDFs processed with translator and OCR.

NLPIR 2024, Okayama, Japan

ACM ISBN 979-8-4007-1738-3/24/12
https://doi.org/10.1145/3711542.3711576

CCS Concepts

• **Information systems** → *Information retrieval*; *Document representation*;

Keywords

Natural Language Processing, Railway Requirement Management, Tender Unstructured Documentation, Multilingual Requirement Classification

ACM Reference Format:
Nocente Arianna, Risi Riccardo, Pannocchia Gabriele, and Rossetti Giulio. 2024. From Unstructured Documents to Annotated Information: An Optimized Pipeline to Process Industrial Requirements. In *2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2024), December 13–15, 2024, Okayama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3711542.3711576

1 Introduction

In the realm of industry, *requirements management* is an area warranting meticulous investigation. Implementing a structured approach when data size and format are non-trivial is a crucial success factor [10]. Discerning between "requirement" and "information" in unstructured documents is essential, strongly affecting tendering and subsequent implementation phases. Requirements undergo periodic updates, often influenced by national regulations and international standards, directly impacting implementation in specific nations. Prior research has leveraged Natural Language Processing technologies in Requirement Engineering (NLPRE), overcoming challenges such as the classification of non-functional requirements (NFRs) with supervised [13] or semi-supervised approach [5], as well as their application in specific domains such as healthcare [18] and automotive [7]. Additionally, NLP tecniques have been used for the categorization of requirements using semantic NLP techniques [15], and for identifying inconsistencies or redundancies within requirements using statistical methods [8]. Furthermore, NLP in

requirements management has been investigated within industrial environments for nearly four decades, contributing to around 7% of all related academic publications [23]. Nonetheless, most of these methods are restricted in their practicality for worldwide industrial use due to their dependence on monolingual datasets or highly structured documentation.

Our contribution addresses these gaps by proposing an optimized pipeline that integrates NLP and Optical Character Recognition (OCR) to support multilingualism and to manage the unstructured nature of industrial documentation, including scanned PDFs with tables and images through effective pre-processing of input data. Furthermore, we tackle the problem of data inconsistency and lack of transparency—common industrial settings—by offering robust classification and the ability to dynamically update, refine, and adapt as new data enters the pipeline. Since the OCR stage could introduce errors and adversely affect the quality of the results by causing misclassifications and inaccuracies [11], a customized interface has been configured for the intermediate steps of text extraction and translation to perform checks. Our semi-supervised method is designed to give users complete control over the process, providing the flexibility to extract text from PDF files and choose the desired granularity for classifying textual items, ranging from sentences and paragraphs to entire pages. During this step, the system issues a warning to the user if the granularity they have chosen will create issues with the text parsing, highlighting the faulty portions of the document. We integrate Optical Character Recognition (OCR) technology to read text embedded in images and tables, allowing for the parsing of scanned PDF files that cannot be processed by conventional PDF extraction libraries. In the subsequent step, we propose two distinct methods for classifying the extracted data into two categories: "requirements" and "information". The first method utilizes a metaclassifier, which is an ensemble of machine learning models, and requires an initial translation of the documents into English. The second method leverages BERT, a multilingual large language model. Both approaches demonstrate comparable performance; however, while the metaclassifier exhibits greater robustness, BERT excels in scenarios where the classification patterns are more nuanced.

2 Method

The proposed pipeline[1]'s primary challenge (Figure 1) is managing and curating multilingual, structured and unstructured data (including PDF scans containing tables and images), thus enabling their classification (either binary and/or multi-class).

As a first step, the unstructured textual data as input - often available as PDF files - are cleaned and transformed into a set of records. A semi-automated fine-tuned OCR approach has been integrated to identify the optimal coordinate systems of text bounding boxes automatically. In those rare instances where bounding the computed box areas overlap, potentially affecting result quality, the devised approach generates a graphical representation for a manual disambiguation intervention.

Once inconsistencies are fixed, documents are ready for the phases of text extraction (using OCR) and automatic translation (leveraging Google Translate API). At the end of such pre-processing, the final, cleaned dataset is composed of set of paragraphs, images, and tables.

Lastly, our pipeline proceeds with the classification phase that occurs by leveraging two alternative versions of the input data: (i) the one translated into English (fed to a metaclassifier based on a majority voting system) and (ii), the non-original, multilingual, dataset (fed to a BERT model).

Data Preparation. Requirements documentation is provided as heterogeneously formatted, non standardized, PDF documents containing text, images, and tables. The objective of this phase is to enable content segmentation: each document is fragmented into textual items, each of which will serve as a minimal input unit in all subsequent analysis and manipulation (e.g., translation).

The content of the PDFs is processed using PDFMiner [2] and Tesseract [17, 19, 22].

Instead of simply detecting all the text in the PDFs, we allow the users to choose the desired granularity for the textual items that they need to classify – such as word-level or paragraph-level. Moreover, our approach is capable of detecting text contained in tables and images. In the first step, PDFMiner is leveraged to detect text with a WYSIWYG (What You See Is What You Get) approach, where a visual interface guides the users in the tuning of the parameters that control the granularity of the bounding boxes built around different chunks of text. Through this process, users can specify whether they desire to classify entire pages, paragraphs, or simple sentences. PDFMiner builds bounding boxes around the textual items, and returns a list consisting of chunks of text and the coordinates of the boxes delimiting them. These coordinates are needed as they allow us to keep track of the relative position and order of the text within the document. Subsequently, Tesseract is used to detect the text contained in the images - thus creating a list of textual chunks along with the coordinates of the bounding boxes encapsulating them. Finally, the two lists are merged and the items sorted using their coordinates, so that the order of the text within the pages is preserved.

Translation. Considering the need for a flexible translation engine without restrictions on character limits, we integrate the Google Translate service (supporting 106 languages) within the proposed pipeline. Google Translate API is chosen after comparison with Argos Translate and other models from Hugging Face due to the limits on the maximum number of translatable tokens in the open source versions and after comparing their performances (bleu, meteor, NIST, chrF, precision, recall, F1) on a set of target measures on domain-dependent data. The translation process targets the textual entities within the predefined bounding boxes. After this stage, the users have the opportunity to download the translated PDF file. Such a document is constructed by replacing the original text with its translation in the bounding boxes that were detected in the previous phase, and the font size of each text chunk is dynamically computed for it to fit within its box.

Semi-Supervised Item Identification. Technical requirements must be segmented into discrete items to facilitate an analytical pipeline, cloud storage, further processing, and eventual classification. Typically, the text is segmented manually using intuitive criteria like line breaks or punctuation marks - which is the method

[1]Code available: https://rb.gy/nvoj0t

[2]PDFMiner: https://github.com/pdfminer/pdfminer.six

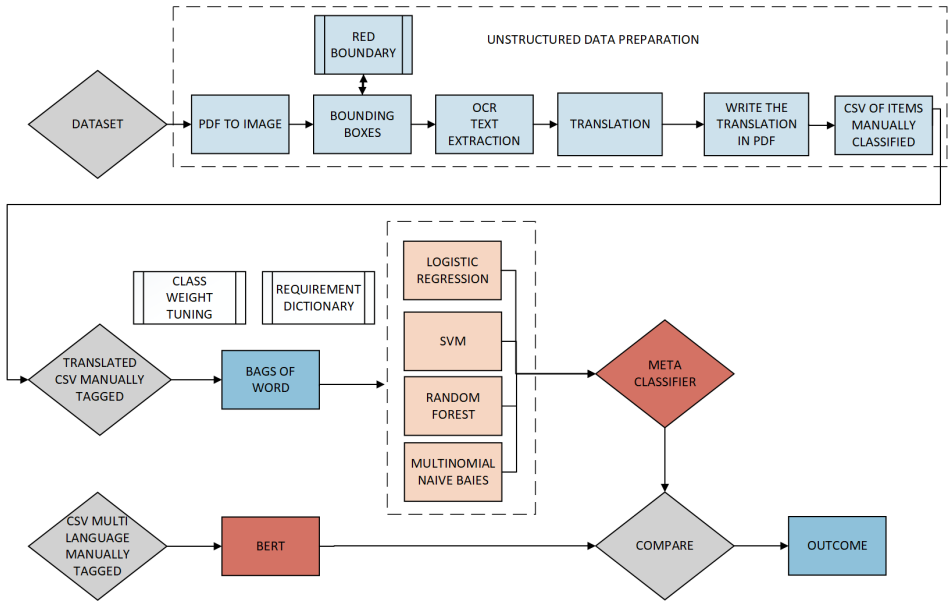


Figure 1: Method Pipeline includes data preparation of unstructured dataset using OCR and correction of overlaps with graphical interface that shows red boundaries, translation of text into English, metaclassifier with majority voting using monolingual classification models and comparison with BERT multilingual.

used to create the reference dataset for validation in the railway case study. Our work assumes that the segments intended for classification align with the document areas identified by the bounding boxes, ensuring that the identified segments serve as the primary units for classification, maintaining analytical consistency and traceability.

Classification. A dual-line pipeline is designed for annotating identified textual items, offering flexibility based on computational resources and task requirements: one branch uses a majority voting metaclassifier on translated texts, and the other employs multilingual BERT on the original, non-translated texts. Both branches of the pipeline use an aligned dataset, divided into training (70%), development (15%), and testing (15%) sets via stratified sampling. The first branch uses a metaclassifier, which combines the predictions of four machine learning algorithms - Logistic Regression [16], Support Vector Machine (SVM) [6], Multinomial Naive Bayes [12], and Random Forest [4] - through majority voting. These classifiers are trained on Bag of Words (BoW) representations, where text is converted into word frequency vectors. BoW is chosen over modern techniques like word embeddings due to its simplicity and low computational overhead, making this approach suitable for environments without access to GPUs. The second branch employs multilingual BERT, which processes the original, non-translated texts, leveraging its pre-training across 104 languages for strong cross-lingual generalization. BERT captures complex linguistic patterns and context, though it requires greater computational resources, including GPUs, to handle its deep learning architecture. To optimize performance, we cap the sequence length at 100 tokens during training. While both methods achieve comparable classification performance, the metaclassifier offers a robust, resource-efficient option, whereas BERT excels in tasks requiring more nuanced contextual understanding. DataLoaders are created for each dataset to manage data batches during training, development, and testing. Evaluation for both pipelines involves comparing precision, recall, F1-score, and support. The majority voting metaclassifier's performance is directly compared with the BERT classifier's results.

3 Experiments

The proposed pipeline has been applied to two distinct case studies: (i) the first involves binary classification with multilingual and unstructured industrial data that requires processing in compliance with safety constraints, and (ii) the second aims to validate the pipeline's applicability to other contexts. The second case study has been selected as a result of the peculiarities observed in publicly available datasets (and their ground truth annotations) specific to requirement identification and classification, which do not meet the ones of our primary case study. For instance, the aeroBERT dataset [21] composed of 310 English sentences classified into design, functional, and performance requirements. However, due to the OCR technology used to process unstructured data and to split the document into items to be classified, sentence-level granularity is not guaranteed, hence a comparison would not be possible. The same rationale applies to the exclusion of other open datasets [2]. While these datasets are commonly used to benchmark NLP pipelines, this paper focuses on developing a pipeline that integrates multilingual inputs and preprocessing techniques for unstructured documents, a domain that remains largely underexplored in industrial requirement applications. Other datasets like PURE [9], comprising English-language structured and unstructured public documents collected from the web could have been employed in this study. However, the lack of reliable annotated ground truth makes them unfitting to evaluate the proposed pipeline. Consequently, we chose to expand the pipeline's scope beyond requirement management

to broader contexts, such as classifying open-access newspaper articles—a domain that is often characterized by unstructured multilingual data and the necessity of leveraging historical knowledge.

Railway Case Study. The first case study uses an industrial dataset containing the text of railway tenders to design Radio Block Center compliant with the European Rail Traffic Management System (E.R.T.M.S.) standards for Denmark, India, and Italy. The original PDFs - along with the related manually processed and annotated CSV files - were provided by Alstom [3]. The pipeline introduced in Section 2 was originally developed by the authors to design a continuous solution for the data owners. A mapping of the original tag labels (e.g., Requirements, Information, Technical, Non-Technical, Management, etc.) was performed by domain experts, thus consolidating a binary label system: requirement or information. After consolidation, the Alstom dataset has 12204 items: 8547 tagged as *requirement* and 3657 as *information*. In the railway sector safety domain, it is crucial to minimize false positives, since misclassifying requirements as information means losing these important specifications that instead have to be implemented mandatorily. To support such a goal, two additional steps have been performed during the classification phase: data augmentation through a dictionary of context-dependent words and class weight tuning. Both additional steps (being highly context-dependent) are deemed optional for the proposed pipelines. However, our analyses underlined that neither has a relevant impact on the classification performances. In Table 1, we report the evaluation obtained by the designed classification pipelines.

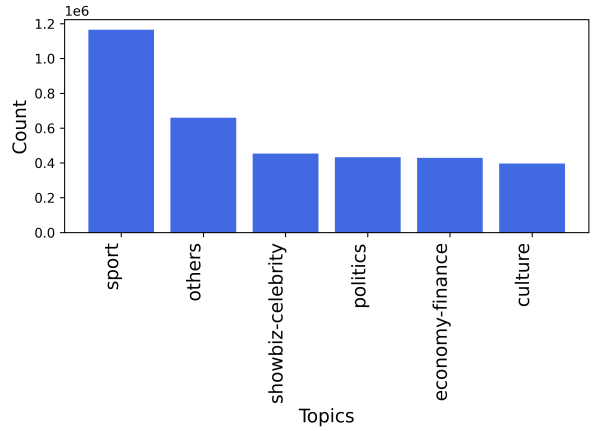
Table 1: Evaluation of Monolingual Text Classifier and Meta(classifier)/BERT on Alstom railway dataset.

Model	Accuracy	Class	Precision	Recall	F1
LR	0.76	Information	0.59	0.77	0.66
		Requirement	0.88	0.76	0.82
SVM	0.76	Information	0.59	0.72	0.65
		Requirement	0.86	0.78	0.82
RF	0.80	Information	0.69	0.64	0.66
		Requirement	0.84	0.87	0.86
MNB	0.77	Information	0.64	0.58	0.61
		Requirement	0.82	0.85	0.84
META	0.77	Information	0.61	0.75	0.67
		Requirement	0.88	0.78	0.83
BERT	0.83	Information	0.78	0.61	0.68
		Requirement	0.84	0.92	0.88

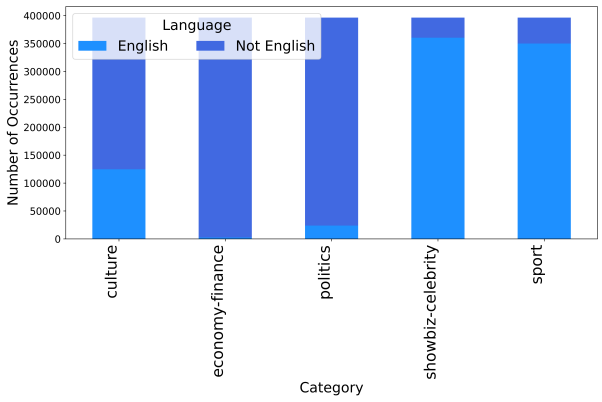
When assessing accuracy, all algorithms yield comparable results, hovering around 76%. However, the precision analysis highlights that RF outperforms others, exhibiting the highest precision for both classes at 69% and 84%, respectively. Similarly, scrutinizing recall metrics, RF achieves notable values for both classes, registering values of 64% and 87%. Furthermore, considering the F1-score, which harmonizes precision and recall, RF continues to excel, boasting figures of 66% and 86% for the respective classes. Consequently, based on these comprehensive evaluations, RF emerges as the best-performing model for the task of requirement classification among the ensemble of algorithms considered. This nuanced analysis provides valuable insights for model selection and optimization strategies in similar classification tasks within the domain. Building a metaclassifier based on majority voting that leverages all the tested classifiers offers several advantages, in terms of robustness, generalization, class imbalance mitigation, and overfitting reduction - particularly when dealing with diverse datasets or classification objectives. Table 1 also reports a comparison of the metaclassifier with the multilingual BERT. The metaclassifier achieves a balanced precision-recall trade-off across both classes, with precision values of 0.61 and 0.88 for class 0 and 1, respectively, and recall values of 0.75 and 0.78. This indicates a more equitable performance across different evaluation metrics. Moreover, the metaclassifier demonstrates consistent performance across different evaluation metrics, as evidenced by the similarity between macro and weighted averages. This suggests a stable and reliable classification performance across the dataset. However, while the metaclassifier provides competitive performance, BERT outperforms it in terms of precision for class 0 and F1-score for class 1. This highlights BERT's strength in capturing intricate linguistic patterns and nuances within the text data, which is particularly beneficial for more challenging classification tasks. Moreover, while the metaclassifier offers robust and balanced performance across various metrics, BERT captures subtle linguistic features, making it a preferred choice for tasks requiring fine-grained classification or handling complex textual data. Ultimately, the choice between the metaclassifier and BERT depends on the specific requirements and priorities of the classification task at hand and the degree of interpretability of the classification decision required by the final user.

Newspaper Case Study. Similarly to what has been done for the railway sector case study, we apply the proposed pipeline to a similar problem: newspaper article topic classification. The adopted public newspaper dataset [14] is composed of 18 million European news articles from 206 media outlets belonging to 27 European countries, published between 2017 and 2021. A considerable fraction of the news articles in the dataset are pre-tagged with their topics by the dataset curators (as inferred from available metadata). These features make it a viable dataset to reapply the pipeline systematically, overcoming the problem of the data having to be manually annotated by experts. Figure 2(a) shows the original topic distributions within the dataset. To mitigate issues arising from the highly skewed distribution of topic frequencies, the analysis is limited to five topics: culture, economy-finance, politics, showbiz-celebrity, and sport. This filtering resulted in a derived dataset consisting of 297,314 articles written in 21 languages. Since the purpose of this case study is not to validate the quality of the translation step, but rather, the classification accuracy - and given the predominant use of the English language in all categories (as underlined by Figure 2(b)) - we focus our analysis on English-based articles only, with balanced class distribution. This latter selection additionally restricts our dataset to 12.545 articles, with about 2.500 articles for each of the five tagged categories.

[3] https://www.alstom.com/ - Third-party dataset access not provided due to industrial confidentiality



(a) Topics Distribution



(b) Languages Topic Distribution

Figure 2: Newspaper dataset: (a) Topics Distribution and (b) Languages Topic Distribution: for each category of the target variable is graphed as the number of articles in English and in all other languages.

As we have done for the industrial case, we evaluate the designed classification pipelines for the newspaper's domain in the following section. From Table 2, we can observe that each algorithm yields comparable accuracy results, with all hovering around 88% on average. However, precision analysis shows that LR excels, particularly in Economy-Finance (96%) and Sport (98%) with SVM and MNB also performing well in those classes. In terms of recall, LR and SVM are effective in Politics (96%), while RF is notable in Economy-Finance and Sport (95% and 97%, respectively) F1-score analysis highlights LR as a top performer, especially in Politics and Sports, with SVM and RF also showing robust results. Overall, LR is a strong contender for multiclass classification, though SVM and RF are also viable options depending on the dataset and specific needs.

Also from Table 2, we can observe that the metaclassifier, constructed through a majority voting system from various base classifiers, exhibits commendable performance across different evaluation metrics when applied to the newspaper dataset. Notably, it achieves balanced precision and recall values for most classes, with precision ranging from 0.86 to 0.99 and recall ranging from 0.90 to 0.97. This indicates a robust precision-recall trade-off, suggesting effective classification across diverse topics. Moreover, the metaclassifier demonstrates consistency in performance across various evaluation metrics, as indicated by the similarity between macro and weighted averages. This consistency underscores its stability and reliability in handling multiclass classification tasks within the dataset. Despite its competitive performance, the metaclassifier is surpassed by BERT in certain aspects. Specifically, BERT outperforms the metaclassifier in terms of precision for certain classes, such as Culture and Showbiz-celebrity, and in the F1-score for Politics. This underscores BERT's capability to capture intricate linguistic patterns and nuances within the text data, which is particularly advantageous for complex classification tasks. While the metaclassifier offers robust and balanced performance across various metrics, BERT is a superior choice for tasks requiring fine-grained classification or complex textual data. Ultimately, the selection between the metaclassifier and BERT depends on the specific requirements and priorities of the classification task, including the level of interpretability desired in the classification decisions.

4 Conclusion and Discussion

This study introduces and evaluates a robust pipeline tailored for the continuous consolidation and annotation of multilingual industrial process data, often presented in non-standardized formats. The purpose of this study is not to introduce a novel algorithm or to perform an exhaustive comparison with existing models—which have already been thoroughly explored in systematic reviews [3], [20] and sector-specific research where more than 20 classification pipelines have been compared [2]. Instead, our work focuses on harmonizing recent advancements in NLP and computer vision to enhance practical value by efficiently processing both structured and unstructured tender documents, linking requirements and their consequent implementations directly to their source documents, thus enabling continuous change impact analysis (CIA). The structured nature of this pipeline further facilitates early-stage bid similarity analysis, extending prior research carried out at Alstom regarding the correlation between requirements and software implementations[1], which had the gap of starting directly from the requirements already manually identified. Our pipeline capitalizes on the strengths of BERT for capturing complex linguistic patterns, while the METAclassifier ensures robust interpretability and maintains high performance, allowing for effective classification in multiple languages. A key contribution of this work is the integration of advanced pre-processing capabilities, including the handling of scanned, non-standardized documents characterized by multi-column layouts and embedded tables and images with textual content. These steps are enhanced by sector-specific dictionary injection, custom weight tuning, and semi-supervised OCR for generating accurate official output documents in intermediate steps, available in different languages. In future work, we plan to

Table 2: Evaluation of Monolingual Text Classifier and Meta(classifier)/BERT on newspaper dataset.

Model	Accuracy	Class	Precision	Recall	F1
LR	0.94	Culture	0.88	0.89	0.89
		Economy-finance	0.96	0.96	0.96
		Politics	0.95	0.96	0.95
		Showbiz-celebrity	0.93	0.92	0.92
		Sport	0.98	0.97	0.97
SVM	0.93	Culture	0.85	0.89	0.87
		Economy-finance	0.95	0.97	0.96
		Politics	0.93	0.94	0.94
		Showbiz-celebrity	0.94	0.89	0.92
		Sport	0.96	0.92	0.94
RF	0.93	Culture	0.85	0.88	0.86
		Economy-finance	0.94	0.95	0.95
		Politics	0.93	0.94	0.93
		Showbiz-celebrity	0.93	0.89	0.91
		Sport	0.95	0.96	0.96
MNB	0.93	Culture	0.85	0.88	0.86
		Economy-finance	0.94	0.95	0.95
		Politics	0.93	0.94	0.93
		Showbiz-celebrity	0.92	0.89	0.91
		Sport	0.95	0.96	0.96
META	0.94	Culture	0.86	0.93	0.89
		Economy-finance	0.96	0.96	0.96
		Politics	0.96	0.92	0.94
		Showbiz-celebrity	0.96	0.90	0.93
		Sport	0.99	0.99	0.99
BERT	0.93	Culture	0.92	0.90	0.89
		Economy-finance	0.99	0.96	0.98
		Politics	0.91	0.98	0.94
		Showbiz-celebrity	0.89	0.93	0.91
		Sport	0.95	0.97	0.96

extend the pipeline to other industry-critic tasks by annotating open requirement datasets and developing an easy-to-use visual tool exposing all the functionality of the proposed pipelines to the end users.

Acknowledgments

This work is supported by: the EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research); PNRR-"SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013.

References

[1] Muhammad Abbas, Alessio Ferrari, Anas Shatnawi, Eduard Enoiu, Mehrdad Saadatmand, and Daniel Sundmark. 2023. On the relationship between similar requirements and similar software: A case study in the railway domain. *Requirements Engineering* 28, 1 (2023), 23–47.

[2] Sarmad Bashir, Muhammad Abbas, Mehrdad Saadatmand, Eduard Paul Enoiu, Markus Bohlin, and Pernilla Lindberg. 2023. Requirement or not, that is the question: A case from the railway industry. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 105–121.

[3] Manal Binkhonain and Liping Zhao. 2019. A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Systems with Applications: X* 1 (2019), 100001.

[4] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[5] Agustin Casamayor, Daniela Godoy, and Marcelo Campo. 2010. Identification of non-functional requirements in textual specifications: A semi-supervised learning approach. *Information and Software Technology* 52, 4 (2010), 436–445.

[6] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* 408 (2020), 189–215.

[7] Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, and Peter Solc. 2007. Automated classification of non-functional requirements. *Requirements engineering* 12 (2007), 103–120.

[8] Henning Femmer, Axel Müller, and Sebastian Eder. 2020. Semantic Similarities in Natural Language Requirements. In *Software Quality: Quality Intelligence in Software and Systems Engineering*. Springer, 87–105.

[9] Alessio Ferrari, Giorgio Oronzo Spagnolo, and Stefania Gnesi. 2017. Pure: A dataset of public requirements documents. In *2017 IEEE 25th international requirements engineering conference (RE)*. IEEE, 502–505.

[10] Parisa Ghazi and Martin Glinz. 2017. Challenges of working with artifacts in requirements engineering and software engineering. *Requirements Engineering* 22 (2017), 359–385.

[11] Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. 2021. Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. *Proceedings http://ceur-ws. org ISSN* 1613 (2021), 0073.

[12] Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2005. Multinomial naive bayes for text categorization revisited. In *Advances in Artificial Intelligence*. Springer, 488–499.

[13] Zijad Kurtanović and Walid Maalej. 2017. Automatically classifying functional and non-functional requirements using supervised machine learning. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*. Ieee, 490–495.

[14] Virginia Morini, Lorenzo Bellomo, Giulio Rossetti, Dino Pedreschi, and Paolo Ferragina. 2023. *European Multilingual News Articles Dataset with Topic Annotation.* https://doi.org/10.5281/zenodo.10397400

[15] Thomas Moser, Dietmar Winkler, Matthias Heindl, and Stefan Biffl. 2011. Requirements management with semantic technology: An empirical study on automated requirements categorization and conflict analysis. In *Advanced Information Systems Engineering*. Springer, 3–17.

[16] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research* 96, 1 (2002), 3–14.

[17] Faisal Shafait and Ray Smith. 2010. Table detection in heterogeneous documents.. In *Document Analysis Systems* (2010-07-07), David S. Doermann, Venu Govindaraju, Daniel P. Lopresti, and Premkumar Natarajan (Eds.). ACM, 65–72.

[18] John Slankas and Laurie Williams. 2013. Automated extraction of non-functional requirements in available documentation. In *2013 1st International workshop on natural language analysis in software engineering (NaturaLiSE)*. IEEE, 9–16.

[19] Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR.. In *International Workshop on Multilingual OCR* (2009-07-25), Venu Govindaraju, Premkumar Natarajan, Santanu Chaudhury, and Daniel P. Lopresti (Eds.). ACM, 1–8.

[20] Riad Sonbol, Ghaida Rebdawi, and Nada Ghneim. 2022. The use of nlp-based text representation techniques to support requirement engineering tasks: A systematic mapping review. *Ieee Access* 10 (2022), 62811–62830.

[21] Archana Tikayat Ray, Bjorn F Cole, Olivia J Pinon Fischer, Ryan T White, and Dimitri N Mavris. 2023. aerobert-classifier: Classification of aerospace requirements using bert. *Aerospace* 10, 3 (2023), 279.

[22] Ranjith Unnikrishnan and Ray Smith. 2009. Combined Orientation and Script Detection using the Tesseract OCR Engine. In *International Workshop on Multilingual OCR*, Venu Govindaraju, Premkumar Natarajan, Santanu Chaudhury, and Daniel P. Lopresti (Eds.). ACM, 1–7.

[23] Liping Zhao, Waad Alhoshan, Alessio Ferrari, Keletso J Letsholo, Muideen A Ajagbe, Erol-Valeriu Chioasca, and Riza T Batista-Navarro. 2021. Natural language processing for requirements engineering: A systematic mapping study. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–41.