



REMOAC: A Retroactive Explainable Method for OCR Anomalies Correction in Legal Domain

Roberto ABBRUZZESE ^{a,b}¹,

Domenico ALFANO ^{a,b}² and Andrea LOMبارDI ^a³

^a*Eustema S.p.A., Research and Development Centre, Napoli*


^b*Department of Management & Innovation Systems, University of Salerno, Fisciano*


Abstract. Efficient OCR Anomaly Detection and Correction is essential in the legal domain, as it significantly enhances the ability of legal professionals to extract accurate information from documents. This paper presents a novel approach called REMOAC, that improves the performance of a legal text classifier through OCR anomaly detection and correction in legal documents, by actively using state-of-the-art models and explainability techniques. Explainability is a key aspect of our approach, both because we provide transparency and comprehensibility in the OCR anomaly correction process and, more importantly, because we actively use it to improve classification performance.


Keywords. Explainable AI, Anomaly Detection, Natural Language Processing, Legal Documents, Transformers,

1. Introduction

In the domain of legal professions, accurate and efficient OCR anomaly detection and correction is indispensable. Legal professionals frequently work with vast amounts of digitized textual data, much of which is processed using OCR. However, OCR processes may introduce anomalies, challenging precise interpretation and extraction. The accuracy of extracted legal information is pivotal for legal analysis, research, and decision-making processes. The first focus of this work addresses the OCR anomaly detection task, discerning whether a given sentence contains OCR anomalies or not. Furthermore, the central objective of this work relies on OCR anomaly correction in the specific context of Legal Text Classification. We propose **REMOAC**, a method that relies on explainability techniques to identify issues within a sentence that then apply appropriate corrections. In particular, our approach focuses the partial correction of tokens by examining and correcting only those tokens that, by presenting errors, could degrade the performance of the classifier. This approach enhances a Legal Text Classifier's capability to generate accurate results even in real-world scenarios.

¹  <https://orcid.org/0000-0002-8621-6210>

²  <https://orcid.org/0009-0008-7865-0544>

³  <https://orcid.org/0009-0004-5508-5766>

2. Related work

Previous research has used anomaly detection to detect and correct anomalies in text [1,2]. These are frequently inadvertent errors that arise as a result of data transfer, such as from audio to text, image to text, or one language to another. Establishing whether a given sentence in a text contains anomalies or errors and determining its location inside the phrase is the first step in the pipeline for correcting OCR anomalies. Thus, the task of analyzing sentences in a text to determine whether they include OCR anomalies and to identify their places is known as OCR anomaly detection. However, this is not always an easy task [3] because these problems include repeated or missing words, which make the work much more difficult. Additionally, there are instances where anomalies are not anomalies, which can occur for a variety of reasons, including the existence of various style manuals in publications [4]. Once it has been shown that a sentence contains anomalies and a correction has been offered, the anomalies can be categorized. Over the years, a number of classification frameworks with various - and largely random - designs have been created [5].

In general, OCR anomaly detection tasks and NLP in general underwent a revolution with the development of BERT (Bidirectional Encoder Representations from Transformers) [6]. Contextual embeddings, which are built based on the context in which a word appears, were introduced by them [7]. In contrast to the traditional left-to-right pre-training, BERT is intended to be utilized for bidirectional pre-training. As a result, it can be used for numerous NLP tasks. By outperforming predecessors like ELMo [8] or Flair [9], it advanced the state of the art of pre-training and became the de-facto cutting-edge solution.

3. Dataset

Italian law firms and legal practitioners can benefit greatly from extracting information from legal documents such as those of the Court of Cassation (*Corte di Cassazione*), which is the highest court of appeal in Italy and sits at the apex of ordinary jurisdiction. Our dataset consists of 730 documents of the Court of Cassation (520 for training and 210 are for testing) acquired from various sources (*ItalggiureWeb*, *Ricerca Giuridica* and *La Legge per tutti*) to maximize the diversity of the data and obtain representative samples. The raw texts were organized into individual lines for each document so that the model input was a single sentence. Specifically, the training set contains about 136,000 sentences, while the test set contains 60,000 sentences; both of them are label balanced.

3.1. Dataset Perturbation

After obtaining the structured dataset, an identical replica was created. Each individual sentence was subjected to perturbation, mirroring the anomalies that commonly manifest during the parsing of texts by an OCR system. This perturbation process was useful to simulate the real scenarios in which text legal professionals frequently work with digitized textual documents, much of which is processed using OCR. It introduced variations such as character recognition anomalies, formatting discrepancies, and occasional misinterpretations that closely mimic the challenges encountered in OCR applications.

The degree of perturbation applied to the sentences is minimal; therefore, only 10% of the characters for each sentence were disturbed [10]. This choice is motivated by the fact that training a model to detect a slight degree of perturbation makes it capable of detecting all types of disturbance. Hence, the dataset comprises lines of text, and it features two distinct types of labels assigned to each line: "1" for perturbed lines and "0" for non-perturbed lines.

4. Proposed Method

This section introduces our innovative approach for OCR anomalies detection and correction for improving legal text classifier.

4.1. OCR Anomaly Detection

The first step of our approach is to classify a sentence to detect whether they are perturbed or not. Let $D = \{l_1, l_2, \dots, l_K\}$ be the dataset consisting of K sentences. Each sentence is associated with a classification label y that can be 0 or 1, respectively representing non-perturbed and perturbed lines. The model employed in this work is ITALIAN-LEGAL-BERT [11], an Italian Bert model pre-trained with an additional 4 epochs on 3.7 GB of text from the Italian National Jurisprudential Archive. This model's parameters are specifically trained on legal terminology, so it is the best candidate for the task we address in this research. As the core of anomaly detection, the training of the classification model was a starting point. After 8 epochs of fine-tuning on our dataset, the model performed with a **0.97 F1 Score** on the test set. The remarkable outcome underscores the robustness and the model's robustness and classification capability, which minimizes false positives and negatives, making it reliable for OCR anomaly detection.

4.2. OCR Anomaly Correction

The strategy put forward intends to enhance the performance of a Legal Text Classifier by spotting and fixing any anomalies introduced by an OCR module during page scanning, encouraging the proactive adoption of explainable methods. The premise is grounded in the idea that if a phrase has been "perturbed" by OCR anomalies and as a result the Legal Text Classifier is unable to accurately predict the class, if one can rectify that token, the prediction might correct itself. The approach, which is broken down into two phases as shown in Figure 1, begins with a preliminary phase that must be performed just once. In the first step of the active phase, a sentence is classified as potentially "perturbed" or not. When perturbation is detected, we extract the most likely perturbed token, compare it with potential replacements and, upon a similarity threshold, substitute it in the text. The cleaned sentence then undergoes to main classifier. Below, are the details of the two phases that make up REMOAC:

1. **Preliminary Phase:** In the preliminary phase, the previously trained Legal Text Classifier, the subject of the improvement, is analyzed by applying SHAP (SHapley Additive exPlanations) [12] a unified framework for interpreting predictions in order to identify feature importance measures. Specifically, for the labels of interest to be predicted, we extract the first N Words (NW) that, on average, sig-

nificantly influenced the Legal Text Classifier on the specific prediction. This is achieved by computing shape values for all items in the training set. In doing so, the first N most representative words are extracted for each label analyzed. The extracted unique words will go to build our BOMIW (Bag of Most Important Words).

2. **Active Phase: Step 1 - MIEW from OCR Detection:** In the first step the sentence is classified by the OCR Anomaly Detection model. In case the sentence is detected as perturbed, the shape values are calculated, and the word most likely to have influenced the choice of error class prediction, called Most Important Error Word (MIEW), is identified. **Step 2 – Similarity Computation:** In the second step, the MIEW is compared with all instances of the BOMIW through edit distance measurement based on the Longest Contiguous matching Subsequence (LCS) [13] and Levenshtein Distance [14]. Once the word with the highest probability of similarity is selected, if that probability exceeds the set threshold it is identified as a "clean" word. **Step 3 – Word Cleaner:** If in the second step the "clean word" is identified, in the third step the cleaner module replaces all instances of the MIEW with the "clean" word. Eventually, the cleaned sentence is given as input to the Legal Text Classifier for prediction.

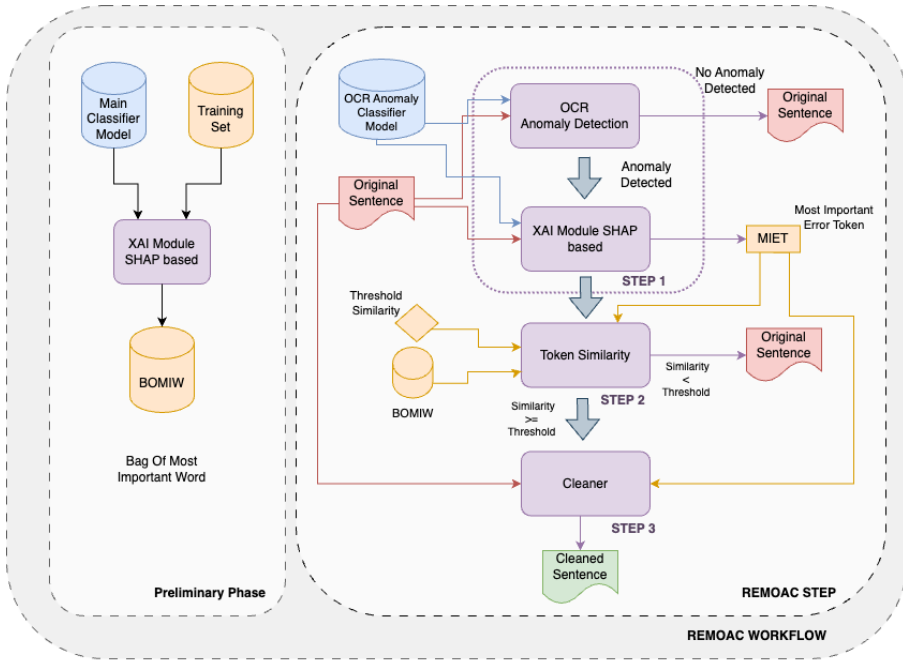


Figure 1. REMOAC Workflow

5. Experimental Results

To evaluate the approach proposed with REMOAC, a use case of text classification in the legal field was identified, more precisely a model that sequentially identifies and clas-

sifies breakpoint sentences, thereby determining the beginning of new sections within the text. The identified classifier was tested with a training set of 68524 samples not perturbed and tested on a fully perturbed dataset of 30298 sentences simulating OCR anomalies. Based on these results, the REMOAC approach was tested considering the same test set as the main classifier. For the preliminary phase, we took into consideration the main classifier and built a 10 words BOMIW. In the active phase, it's essential to note that the test set is entirely perturbed, featuring labels exclusively set to 1. Consequently, this scenario's only viable performance metric is precision, yielding a precision result of 0.963. Various experiments were carried out considering a range of different thresholds for the calculation of similarity specifically: [0.65, 0.70, 0.75, 0.80, 0.85] and the optimal threshold was determined to be 0.80. Also, different edit distance metrics were considered, such as LCS [13] and Levenshtein distance [14], but there are no significant differences in the choice of the algorithm.

Empirically considering 0.80 the similarity threshold with better results, the Table 1 shows the comparison of the results of the main classifier with and without REMOAC. Eventually, we wanted to perform an additional test by removing the OCR Anomaly Detection prediction (Step 1, OCR Anomaly Detection Module) from the active phase, assuming all sentences in the test set as perturbed, but still using the OCR Anomaly Detection Model to extract the MIEW. The Table 1 shows the results comparison. The performance improvement without the OCR Anomaly Detection module happens because the original model error is not propagated on the final result.

Moreover, REMOAC is completely configurable. Some parameters can be changed to evaluate the best configuration to be adopted. Specifically, it is possible to select which labels to analyze to create the bag of most important words, the number of words (NW) per label to analyze to create the bag of most important words, the algorithm for word similarity and the threshold for similarity between words.

Model	Balanced Accuracy	F1 Binary	F1 Multi - Macro avg
Legal Text Classifier without REMOAC	0.82	0.77	0.70
LTC - REMOAC w/ OCR Anomaly Detection prediction	0.94	0.93	0.91
LTC - REMOAC w/o OCR Anomaly Detection prediction	0.94	0.94	0.91

Table 1. Result comparison

6. Conclusion

In this research, we tackle the critical challenge of Optical Character Recognition Anomaly Detection and Correction within the legal text classification domain. Leveraging state-of-the-art XAI methods, we propose REMOAC, a Retroactive Explainable Method for OCR Anomaly Correction designed to address the multifaceted challenges posed by OCR anomalies within this specialized domain, by being concerned with not correcting all the input text but only the tokens that might affect the classification model. We show that our method can improve the performance of the underlying task. The intuition arises precisely in using the SHAP values of the text classifier to construct a limited and contextual bag of words that proposes a reduced but consistent number of substitutions for the perturbed tokens. One promising avenue for future investigation involves the comprehensive evaluation of our approach in different linguistic contexts and in different NLP tasks such as Named Entity Recognition, Topic Detection and Keyword Extraction.

Extending the evaluation to encompass languages beyond Italian and to other domain would shed light on the method's language-agnostic potential.

7. Acknowledgments

This work was completed with the generous resources provided by Eustema S.p.A. Research and Development Centre of Naples and by the Department of Management & Innovation Systems of University of Salerno.

References

- [1] Eskin E. Detecting Errors within a Corpus using Anomaly Detection. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics; 2000. .
- [2] Samanta P, Chaudhuri BB. A simple real-word error detection and correction using local word bi-gram and trigram. In: Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP); 2013. .
- [3] Chodorow M, Dickinson M, Israel R, Tetreault J. Problems in Evaluating Grammatical Error Detection Systems. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee; 2012. .
- [4] Naples C, Sakaguchi K, Post M, Tetreault J. Ground Truth for Grammatical Error Correction Metrics. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Association for Computational Linguistics; 2015. .
- [5] Dahlmeier D, Ng HT, Wu SM. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics; 2013. .
- [6] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2019. .
- [7] Bell S, Yannakoudakis H, Rei M. Context is Key: Grammatical Error Detection with Contextual Word Representations. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics; 2019. .
- [8] Peters ME, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2017. .
- [9] Akbik A, Blythe D, Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics; 2018. .
- [10] Ma E. NLP Augmentation; 2019. <https://github.com/makcedward/nlpaug>.
- [11] Licari D, Comandè G. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management;. .
- [12] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30; 2017. .
- [13] Maier D. The Complexity of Some Problems on Subsequences and Supersequences. J ACM. 1978.
- [14] Miller FP, Vandome AF, McBrewster J. Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau-Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press; 2009.