technologies

IMPACT FACTOR 4.2

CITESCORE 6.7

Article

A Transformer-Based Model for the Automatic Detection of Administrative Burdens in Transposed Legislative Documents

Victor Costa, Mauro Castelli and Pedro Coelho

MDPI

https://doi.org/10.3390/technologies13040134

technologies

MDPI

Article

A Transformer-Based Model for the Automatic Detection of Administrative Burdens in Transposed Legislative Documents

Victor Costa, Mauro Castelli * and Pedro Coelho

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-332 Lisboa, Portugal; vcosta@novaims.unl.pt (V.C.); psc@novaims.unl.pt (P.C.)
* Correspondence: mcastelli@novaims.unl.pt; Tel.: +351-213828610208

Abstract: Legislative impact assessment (LIA) can be defined as the process performed by governments and legislative bodies to evaluate the potential effects of proposed policies or directives before they are implemented. This assessment typically covers various aspects (including economic, social, and environmental impacts) and is designed to ensure that policy proposals are well-founded, transparent, and that potential impacts are thoroughly examined before decisions are made. This process is nowadays performed by human experts and requires a significant amount of time. It is also characterized by some subjectivity that makes it difficult for citizens and companies to perceive the process as a transparent one. Moreover, public administration services responsible for LIA recognize significant difficulties in performing a timely and effective impact assessment exercise due to the lack of human and financial resources. To answer this call, this paper presents an artificial intelligence-based system to automatizing part of the impact assessment process, with the specific objective of detecting administrative burdens from transposed EU legislation. The system is built on a fine-tuned, transformer-based architecture leveraging transfer learning, making it an innovative tool for automating legislative impact assessment. Comprehensive testing on transposed European legislation demonstrated that the system significantly enhances efficiency and accuracy in what has traditionally been a complex and time-consuming task.

Keywords: legislative impact assessment; natural language processing; transformers; deep learning

Academic Editor: Manoj Gupta

Received: 17 February 2025
Revised: 24 March 2025
Accepted: 25 March 2025
Published: 1 April 2025

Citation: Costa, V.; Castelli, M.; Coelho, P. A Transformer-Based Model for the Automatic Detection of Administrative Burdens in Transposed Legislative Documents. *Technologies* **2025**, *13*, 134. https://doi.org/10.3390/technologies13040134

1. Introduction

Legislative impact assessment (LIA) [1] is a critical tool used in the policymaking process to evaluate the potential effects of proposed legislation before its entry into force. This assessment aims to ensure that new laws are effective, efficient, and equitable by estimating their economic, social, and environmental impacts. By providing a formal approach to understand the consequences of legislative actions, LIA helps policymakers make informed decisions that align with public interests and policy objectives. The relevance of LIA has grown significantly in recent decades as governments and international bodies increasingly recognize the importance of evidence-based policymaking [2]. In the European Union (EU), legislative impact assessments are a fundamental part of the legislative process, ensuring that laws contribute to sustainable development, enhance public welfare, and avoid unintended adverse outcomes. Moreover, by evaluating proposed legislation, LIA supports transparency, accountability, and the rational use of public resources [2]. The regulatory framework for LIA within the EU is well-established and robust. The European Commission's Better Regulation Agenda [3] outlines the procedures and standards for conducting

impact assessments. Key components of this framework include the Better Regulation Guidelines and the Better Regulation Toolbox [4], which provide detailed instructions and methodological tools for conducting high-quality assessments. Despite the structured approach and comprehensive guidelines, LIA faces several challenges. One of the main challenges is balancing the need for timely assessments with the complexity of the legislative proposals and the depth of analysis required. In particular, LIA is a time-consuming task that involves a meticulous examination of proposed policies or directives. It involves the collaboration of experts from diverse fields such as economics, law, and the environment. The complexity arises from the need to forecast the effects of a proposed measure, including its economic implications, societal impacts, and environmental repercussions. The thoroughness required in assessing potential outcomes necessitates a comprehensive examination of alternative policy options and potential mitigations. In Portugal, UTAIL (Unidade Técnica de Avaliação de Impacto Legislativo) is the government entity responsible for performing legislative impact assessment. Despite the growing amount of human and economic resources witnessed in UTAIL, the entity recognized significant difficulties in performing a timely and effective impact assessment exercise. Thus, while legislative impact assessment is a fundamental component of modern governance, promoting the formulation of effective and efficient legislation, there is a need to support this process to make it faster and more effective.

To answer this call, in this paper, we address the inefficiency and subjectivity of the LIA process, particularly in identifying administrative burdens within legislative texts. Currently, the process of identifying administrative burdens relies heavily on manual examination by legal experts, which is time-consuming, labor-intensive, and prone to variability based on the subjective interpretations of the reviewers. Administrative burdens, which refer to the compliance costs incurred by individuals, organizations, or businesses to adhere to regulatory obligations, are typically embedded in complex legal language. This makes their detection a highly challenging task, especially when dealing with large volumes of transposed EU legislation. Given the structured yet complex nature of legal texts, a more efficient approach is required to reduce the time and effort involved in analyzing legislation for administrative burdens.

Transformer models [5,6], particularly BERT [7] and its derivatives [8], are well-established in natural language processing. However, their application to LIA and administrative burden detection is novel. Unlike standard NLP tasks such as sentiment analysis or named entity recognition, LIA requires fine-grained contextual understanding of legal texts, where obligations may be implied rather than explicitly stated.

This research goes beyond merely applying pre-existing NLP methods by (1) fine-tuning domain-specific models on transposed EU legislative texts, which had not been previously explored in LIA, (2) developing a noise reduction module to improve data preprocessing for legal text classification, and (3) introducing a human-in-the-loop framework, allowing legal experts to iteratively refine burden detection outputs. While our work leverages established NLP architectures, its contribution lies in adapting these models for a highly specific and underexplored domain [9], ensuring practical applicability to real-world legislative analysis.

Specifically, the objective of this research is to develop an AI-driven tool that can automatically scan and highlight administrative burdens in legislative documents, providing legal experts with a preliminary analysis and allowing them to focus on higher-order tasks in the LIA process. By automating this process, we aim to offer a tool that facilitates quicker and more consistent identification of burdens, improving the overall quality of the regulatory analysis. This study presents a systematic approach to leveraging state-of-the-art NLP models to address this problem, demonstrating their effectiveness in analyzing leg-

islative documents and assisting in the refinement of legislative frameworks. The ultimate goal is to develop a framework that streamlines the LIA process, providing policymakers with objective, data-driven insights that enhance decision-making when drafting or modifying legislation.

Our research yielded remarkable results, showing that artificial intelligence can revolutionize the legislative impact assessment paradigm. The prototype proved highly effective in accurately identifying administrative burdens within legislative texts, providing a rapid impact assessment report. In particular, the proposed system (AI4IA-Artificial Intelligence for Impact Assessment) produced the following impacts:

- Contribution to a more efficient and less time-consuming impact assessment exercise, allowing for faster comparison between legislative proposals or between a proposal and existing legislation;
- Facilitating the analyses of national and transposed EU legislation for the identification of overlapping obligations for businesses or citizens.

We can thus consider this AI-based prototype as a first step in the development of a platform that explores the use of data-driven technologies in support of better regulation.

The manuscript is organized as follows. Section 2 introduces the transposition process of the EU legislation to the internal (of Portugal) legal framework. Section 3 describes the proposed system and the data used to train the underlying deep learning models. Section 4 lists the experimental settings and and analyzes the experimental results. Finally, Section 5 concludes the paper and suggests future research avenues.

2. Transposition of Legislation: Legal Significance and Process

The legislative impact assessment process must be preceded by a clear description and understanding of all stages of the legislative process. The transposition of directives, being a process that comprises the interconnection between the European and national levels, has specificities that must be considered.

Transposition is a critical process within the European Union's legal framework, wherein EU directives are incorporated into the national legal systems of the member states. Unlike EU regulations, which have direct effect and are automatically applicable across all member states, EU directives set goals or objectives that each member state must achieve but allow the national authorities flexibility in determining how to implement them. This flexibility is essential because member states have different legal traditions, administrative structures, and regulatory environments. The process of transposition involves translating these EU directives into national laws that reflect the specific legal and administrative contexts of each country while ensuring compliance with the objectives set out by the directive.

The significance of transposition lies in its role as a legal mechanism that ensures the uniform application of EU law across diverse legal systems. By transposing directives, member states are bound to meet the minimum standards or goals set by the EU while retaining the ability to adapt the measures in a way that suits their domestic legal frameworks. Failure to transpose directives correctly or within the specified timeframes can result in infringement procedures initiated by the European Commission, potentially leading to financial penalties or legal action before the Court of Justice of the European Union.

The transposition process involves several stages. First, the directive is examined by national legislative bodies to determine how its provisions align with existing national laws. This examination is crucial to identifying the legal areas where changes, updates, or new laws are required to meet the directive's objectives. During this phase, legal experts analyze whether current national legislation already satisfies the directive's requirements or if additional legislation is necessary. Once the need for legislative changes is established,

the next step is to draft the necessary national laws, ensuring that the provisions of the EU directive are accurately reflected in domestic law. This drafting process often requires close collaboration between national lawmakers, legal experts, and relevant government agencies to ensure that the law's intent is preserved while fitting within the country's existing legal structure.

After the draft law is prepared, it must be approved by the appropriate legislative body—such as the national parliament. In some cases, public consultations may also be held to gather input from stakeholders, such as industry representatives or civil society organizations, who may be affected by the new legislation. Once approved, the transposed law enters into force, formally implementing the provisions of the EU directive into national law. In many cases, the directive sets a deadline by which member states must complete the transposition process. Failure to meet this deadline can result in non-compliance with EU law, triggering oversight actions from the European Commission.

It is important to note that while transposition ensures the achievement of common EU goals, the specific legal measures enacted by member states may differ due to variations in national legal systems. This divergence means that, while the overarching objectives of the directive are uniformly applied, the practical implementation may look different in each member state. In some cases, this can lead to discrepancies in the application of EU law, which the European Commission monitors closely to ensure uniformity and adherence to EU standards.

In the case of Portuguese law, the process of transposition follows a similar structure. When an EU directive is issued, the Portuguese legislative body reviews the directive's content to identify any necessary legislative changes. This process typically involves the Assembleia da República (the Portuguese Parliament) and other relevant bodies, which may need to amend or introduce new laws to comply with the directive. Portugal's legal tradition, which emphasizes administrative clarity and adherence to international commitments, ensures that EU directives are transposed in a manner that aligns with national legal standards while meeting the directive's requirements.

In summary, the transposition of EU directives into national law ensures that EU objectives are achieved across member states while allowing flexibility in how those objectives are implemented. This process requires careful translation and adaptation of EU laws into national contexts, ensuring both compliance with EU standards and the preservation of domestic legal integrity.

Figure 1 summarizes the EU legislative process that culminates with the transposition of the EU legislation to the internal legal framework. Throughout this process, the regulatory impact assessment plays a crucial role in supporting the EU legislative process and opens the possibility of a closer intervention by the member states and national impact assessment units.

First, during the initial discussion on the measures to be implemented and the results presented in the EU Commission Regulatory Impact Assessment (EU COM RIA), national impact assessment units may contribute by providing evidence on the impacts in each member state or by providing comments regarding the Commission's conclusion. This contribution became even more relevant with the implementation of the "one in, one out" principle that considers "Adjustment costs" and "Administrative costs". The impact assessment unit of each member state may internally provide information not only considering the internal impacts of the proposed legislation, but also informing the potential impact of the "one in, one out" principle. Any contribution of the national impact assessment units at this stage is particularly time-sensitive and may require the use of relevant resources.

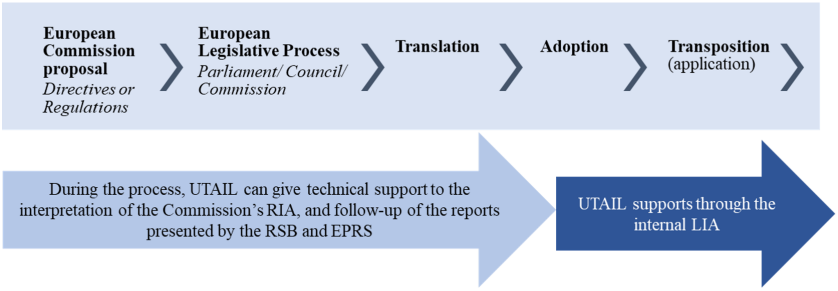

Figure 1. Simplified representation of the EU legislative process and its interaction with UTAIL.

Still, during the initial preparation of a new law and the European legislative process, there might be substantial amendments in the Council or the Parliament that might be the object of an impact assessment. Again, and even in a more time-sensitive situation, the assessments are required to be robust and informative to support negotiation and final decision-making. We point out that the aforementioned process precedes the national impact assessment process, which focuses on the transposed legislation. In the context of this project, we will only focus on the national transposed legislation.

The European Commission started to implement strategies for reducing regulatory burdens in 2002, and, over the years, the approach to burden reduction has been systematized with the REFIT program [10]. REFIT aims to make EU laws simpler, more targeted, and easier to comply with. For this purpose, following the "one in, one out" principle, it offsets any new burden for citizens and businesses resulting from the Commission's proposals by removing an equivalent existing burden in the same policy area. In this way, a new regulation will not add unnecessary regulatory burdens. In the law-making process, the Commission recognizes the importance of impact assessment in providing evidence to inform and support the decision-making process. For this reason, the Commission provided guidelines on how to conduct impact assessments (https://ec.europa.eu/info/sites/default/files/swd2021_305_en.pdf, accessed on 9 January 2025) and a toolbox that provides complementary advice (https://ec.europa.eu/info/sites/default/files/br_toolbox-nov_2021_en_0.pdf, accessed on 9 January 2025). Beyond that, the Joint Research Centre (JRC), the European Commission's science and knowledge service, has developed a tool, SeTA (Semantic Text Analyser) [11], which applies advanced text analysis techniques to large document collections, helping policy analysts to understand the concepts expressed in thousands of documents and to see visually the relationships between these concepts and their development over time [11]. In more detail, SeTA creates language models based on the well-known neural word embedding [12], where a recurrent neural network is fed with patterns of words and can learn the distribution function of the words. The main idea is to learn the relations between terms. SeTA follows the main principle of Word2vec [13], a two-layer neural network that processes text. However, to overcome the main limitation of Word2vec (i.e., it does not capture meaning related to word order), SeTA encoded phrases alongside simple words. The key observation was that while in general English the number of identified phrases would quickly grow, the specialized legal phrasal dictionary is rather limited. To discover these phrases directly from the plain corpus, SeTA's developers relied on sentence dependency parsing, where, for every noun, the complete dependency tree has been calculated. In this way, it was possible to embed phrases and words to encode the differentiated meaning of words and phrases. For instance, the words "artificial" and "intelligence" have very different meaning vectors from the phrase "artificial intelligence". The resulting language model allows for representation of the knowledge base of the Commission without a need for any tailored algorithm. To calculate the vectors, SeTA's developers used FastText rather than Word2Vec because of its higher-quality representations. The neural network learns, in an unsupervised manner, the relations between the

considered terms in a sentence, within a moving window, which means that it learns how the context in which the term is used can lead to its meaning, both in terms of synonyms and of the area of application or context in which it is used. A pilot version of SeTA was fed with a vast set of documents from EUR-Lex and used at the JRC in many policy-related use cases, including impact assessment.

Still focusing on impact assessment, another initiative of the European Commission (EC) that was developed starting in 2013 is the Modelling Inventory and Knowledge Management System (MIDAS) [14], a scientific information system developed and managed by the Competence Centre on Modelling (CC-MOD) of the EC to document the models run by or on behalf of the Commission and their contributions to Commission's impact assessments. Based on the description of the developers, MIDAS provides an integrated set of components for collecting, storing, processing, and sharing metadata on model structure, quality, and transparency, as well as access to model documentation, useful references, and the supported EC's impact assessments. This initiative aims to enhance the transparency of policy-relevant models and the traceability of their use, as well as facilitate the understanding of impact assessments performed by the EC. In this vein, MIDAS follows the guidelines of the Better Regulation agenda by promoting a transparent use of modeling that underpins the evidence base of EU policies [14].

To analyze how models are used to support the work of the EC, the same Commission, in 2019, presented a study [15] that discusses the use of models in European Commission impact assessments. The goal of this analysis was to achieve a better understanding of these models and how they can contribute to the sound use of evidence in support of EU policies. For this purpose, the authors took into consideration a total of 1063 impact assessment tasks carried out in the years 2003 to 2018 and examined the frequency and characteristics of model use by using text mining techniques complemented by manual post-processing. In their research, they relied on MIDAS to trace models used to support policies. The main results of this analysis suggested that models are used in 16% of the total impact assessment tasks. In particular, the authors identified 123 different models that have been developed or run by the EC or by third parties. However, more than half of these models were used only once, thus raising some concerns related to the efficiency of model use and reuse in the EC, as well as on the scope for improved coordination of model-related products and services. This is even more evident considering that the top 10 models were used in 66% of the total number of impact assessment tasks (i.e., the ones relying on models). Additionally, policy areas with the highest number of impact assessment tasks using models are the environment (including climate), internal market, transport, and energy. As a consequence, the authors concluded that further action is needed to use and promote best practices to ensure transparency and accessibility over time of the evidence base in support of impact assessments. This is particularly important for the regulatory impact assessment and the analysis of national and transposed EU legislation, where no models have been identified by the authors of the aforementioned study. In particular, there is no evidence about the possibility of using AI-based techniques to fully support the regulatory impact assessment process. This project aims to answer this call by producing a proof of concept where AI techniques are employed to support human experts in one of the preliminary steps of the law-making process. In this way, this project links the objectives of the "Shaping Europe's digital future" agenda with those of the Better Regulation agenda.

Existing AI-Based Initiatives for Impact Assessment

This section describes the existing initiatives in which AI was proposed in the context of impact assessment of legislation. Despite the potential benefits that AI can bring to public governance and the objective of the EU to foster the use of AI in the decision-making

process [16], the use of AI for the analysis of legislation is still in its infancy. To the best of our knowledge, at the European level, the proposed AI4IA system is the only project where AI was used to support the impact assessment task. In particular, concerning European countries, the existing studies/projects are mainly focused on suggesting strategies to enhance the analytical abilities of the national legislation offices [17]. The final goal is to lead to the adoption of appropriate methodological tools and processes of regulatory impact assessment to improve the countries' overall regulatory performance. In this vein, two projects proposed for Croatia [17] and Germany [18] highlighted the need for a modern and automatic regulatory impact assessment, by pointing out the potentiality of artificial intelligence. Similarly, a report by Deloitte (https://www2.deloitte.com/content/dam/Deloitte/de/Documents/risk/Deloitte-Kostbar-Abschlussbericht-EN.pdf, accessed on 9 January 2025) showed that, in Germany, companies are subject to a considerable number of laws and regulations, with insurance and mechanical engineering companies spending between 4 and 7 percent of their annual personnel and material expenses to comply with the federal relevant regulations. These regulatory expenses can make or break a company's profitability. For this reason, the study highlighted the importance of reducing these costs by adopting NLP techniques that may simplify the existing regulations. In particular, when asked how legislation should be designed to reduce regulatory burdens, companies highlighted the importance of pursuing better regulation through advanced data science methods. However, despite the existence of these studies and initiatives, no further steps were taken toward the implementation of AI-based systems. Karpen [19] described the current methods of regulatory impact assessment at the European and member state levels, and, in both cases, there is no evidence of AI-based systems being used to support this task. Other initiatives highlighting the suitability of AI for improving public governance and regulatory impact assessment exist [20] but, to the best of our knowledge, no country has moved towards the implementation of automatic systems to support this task. This is mainly due to the lack of expertise as well as the difficulty in foreseeing the benefits provided by AI in this area.

3. Methodology and Data

The proposed AI4IA system aims to improve the efficiency of regulatory impact assessment by automating the identification of administrative burdens within legislative documents. In particular, the AI4IA framework must provide the following functionalities:

- Scalable Impact Assessment: the system receives as input existing or proposed directives and generates a comprehensive impact assessment report. In particular, through the advanced capabilities of a natural language processing model, the system must identify specific paragraphs where administrative burdens are present;
- Legislation Comparison: the system must provide a feature to compare two directives. This functionality would enable the evaluation of either two national Portuguese directives or a European directive against its transposed national version. By detecting discrepancies and gold plating (the practice of adding more stringent or extensive requirements to a piece of legislation than what is required by the original directive or regulation), this functionality of the prototype would enhance transparency and facilitates a deeper understanding of legislative variations.
- Human Expert Feedback and Continuous Learning: This feature allows for a continuous feedback mechanism. In particular, human experts retain the ability to correct the system's predictions. This feedback loop would establish a collaborative relationship between AI and human experts. In particular, the system triggers alerts when a sufficient number of human-validated administrative burdens accumulate, signaling

the possibility for model retraining. Thus, this feature of the prototype ensures its adaptability to evolving legislative landscapes.

The system answers the need for understanding text at scale by employing NLP techniques [21]. Concretely, the objective is to alleviate the impact assessment expert of personally reviewing each passage in the transposed legislation and, instead, train an NLP model to identify the existing administrative burdens as a text classification setting. With such a goal in mind, we proposed a "filter model" that would be able to receive as input the text passages of a legislative document and classify each passage as either containing or not an administrative burden—i.e., a binary, text classification task. With such a capability, the system is then capable of speeding up the expert's manual work.

Large, pre-trained language models [22] are considered the state of the art in predictive NLP and, therefore, were explored for building the envisioned model. Unlike traditional recurrent neural networks [23] or convolutional neural networks [24], transformers do not rely on sequential data processing. Instead, they employ a mechanism known as self-attention [25], which allows the model to weigh the importance of different words in a sentence, regardless of their position. This mechanism enables transformers to handle long-range dependencies and capture contextual information more effectively. Section 3.1 describes the architectures considered in this work and Section 3.1 outlines the fine-tuning procedure.

3.1. Considered Transformer-Based Models

The bidirectional encoder representations from transformers (BERT) model [7] is one of the foundational architectures in the family of transformer-based models for natural language processing tasks. BERT revolutionized NLP by applying a bidirectional approach to understanding the context of a word in relation to its surrounding words, as opposed to the unidirectional models used before. BERT's architecture is built on the transformer encoder, a component of the transformer architecture introduced by Vaswani et al. in 2017 [5]. The key innovation in the transformer model is the self-attention mechanism, which allows BERT to weigh the importance of each word in a sentence relative to all other words in the sentence, thus capturing the nuanced relationships between them.

The transformer architecture underlying BERT consists of multiple layers of encoders, each composed of two primary sub-layers: a multi-head self-attention mechanism and a feedforward neural network. The self-attention mechanism enables the model to capture the dependencies between words across different positions in the sentence by assigning attention scores. This is particularly effective for handling complex linguistic structures. Additionally, BERT leverages a bidirectional training mechanism, meaning it considers both the left and right context of a word during training. This bidirectional approach distinguishes BERT from earlier models, which were limited to a unidirectional understanding of context.

BERT is pre-trained on large corpora using two tasks: masked language modeling, where some percentage of the input tokens are masked and the model is trained to predict them, and next sentence prediction, which helps the model understand sentence relationships. After pre-training, BERT can be fine-tuned on specific tasks, such as classification, named entity recognition, or in this case, the detection of administrative burdens in legislative texts. This architecture forms the backbone of transformer-based NLP models like Albertina [26] and BERTimbau [27], both of which build on BERT with modifications tailored to specific use cases and languages.

Albertina is a derivative of the ALBERT model [28], which stands for A Lite BERT. The goal of Albertina is to retain the core strengths of BERT, such as its powerful contextual understanding and bidirectional encoding, while significantly reducing the model

size and improving computational efficiency. Albertina achieves this through several architectural innovations designed to address the limitations of BERT, particularly its resource-intensive nature.

One of the primary modifications in Albertina is parameter sharing across the transformer layers. In BERT, each layer of the transformer has its own set of parameters, which leads to a significant increase in model size as the number of layers grows. Albertina, on the other hand, shares parameters between layers, which greatly reduces the number of parameters without compromising the model's ability to learn complex language patterns. This parameter sharing is applied to both the feedforward networks and the attention mechanisms within the transformer layers.

Another key feature of Albertina is the introduction of factorized embedding parameterization. In BERT, the size of the vocabulary embeddings is tied directly to the hidden layer size, leading to large embedding matrices that consume significant memory. Albertina decouples these dimensions by factorizing the embedding matrix into two smaller matrices, allowing for a more compact representation while still maintaining expressiveness.

Despite these optimizations, Albertina retains BERT's bidirectional training approach and continues to use the same pre-training objectives: masked language modeling and next sentence prediction. However, due to its lightweight design, Albertina is particularly well-suited for applications where computational resources are limited or where fast inference times are required, such as real-time analysis of legislative texts.

Another architecture considered in this research is BERTimbau [27], a BERT-based model specifically designed for Portuguese language tasks. While the core architecture of BERTimbau is identical to BERT in terms of the transformer layers and the self-attention mechanism, the distinguishing feature of BERTimbau lies in the corpus used for pre-training. BERTimbau was trained on a large-scale corpus of Portuguese texts, which enables it to capture the linguistic nuances, idiomatic expressions, and grammatical structures unique to the Portuguese language.

BERTimbau benefits from the same transformer architecture as BERT, with its multi-layer bidirectional encoders that process text by assigning attention scores to words based on their relationships with other words in the sentence. However, its Portuguese-specific pre-training makes it highly effective for NLP tasks within this language domain, such as the detection of administrative burdens in Portuguese legislative documents.

The pre-training process of BERTimbau involved using both formal and informal Portuguese texts, ensuring that the model can handle a variety of registers, from legal texts to conversational language. This wide-ranging pre-training enables BERTimbau to perform well in tasks that require a deep understanding of the language, such as detecting the nuanced and often complex phrasing used in legal and administrative documents.

BERTimbau retains the same masked language modeling and next sentence prediction pre-training tasks as BERT, but because it is fine-tuned on Portuguese-specific datasets, it has a clear advantage over more general models when working with Portuguese texts. The ability to understand the linguistic and cultural context of Portuguese allows BERTimbau to provide more accurate and contextually relevant outputs, making it a powerful tool for analyzing legislative texts in the Portuguese language.

In this research, the core idea was to take advantage of a transfer learning setting [29], where knowledge previously learned by the model from a vast amount of generic data is repurposed and fine-tuned for a specific, downstream task. Recently, the field of NLP has capitalized on transfer learning techniques by using vast amounts of freely available text from the web to yield pre-trained models that significantly improved upon state-of-the-art results for several NLP tasks [30]. Specifically, three pre-trained models were evaluated for the AI4IA system. Two of these were developed by Unicamp (State University

of Campinas, Brazil) in partnership with the University of Waterloo in Canada and are available in the open repository Hugging Face. "BERTimbau Base" and "BERTimbau Large"—as the models are called—leverage the original BERT's architecture [31] and were both pre-trained on a Brazilian Portuguese corpus (namely, the brWaC dataset), which is composed of 2.7 billion tokens, during 1,000,000 steps. The main difference between these two is the size: 'Base' has 110 million parameters, while 'Large' has 335 million parameters. More information can be found in the models' github repository (https://huggingface.co/neuralmind/bert-large-portuguese-cased, accessed on 9 January 2025). The third pre-trained model leveraged for experimental evaluation was the 100 million parameter variant of the "Albertina" family of models. "Albertina 100M PTPT", as it is called, is a language model for European Portuguese. It was trained over a 2.2 billion-token dataset that resulted from gathering openly available corpora of European Portuguese from the following sources:

- OSCAR: the OSCAR dataset includes documents in more than one hundred languages, including Portuguese, and it is widely used in the literature. It is the result of a selection performed over the Common Crawl dataset, crawled from the Web, that only retains pages whose metadata indicate permission to be crawled, performs deduplication, and removes some boilerplate, among other filters. Given that it does not discriminate between the Portuguese variants, extra filtering was performed by retaining only documents whose metadata indicate the Internet country code top-level domain of Portugal. The January 2023 version of OSCAR was used, which is based on the November/December 2022 version of Common Crawl;
- DCEP: the Digital Corpus of the European Parliament is a multilingual corpus including documents in all official EU languages published on the European Parliament's official website. The European Portuguese portion was retained;
- Europarl: the European Parliament Proceedings Parallel Corpus is extracted from the proceedings of the European Parliament from 1996 to 2011. The European Portuguese portion was retained;
- ParlamentoPT: the ParlamentoPT is a dataset obtained by gathering the publicly available documents with the transcription of the debates in the Portuguese Parliament.

The pre-trained model and its variants were developed by a joint team from the University of Lisbon and the University of Porto, and further details can be found in [32]. Larger pre-trained models were avoided due to considerations regarding how the system should work in production. The main goal of this research is to streamline the impact assessment analysis, which by consequence should translate into a reasonably fast inference time. Considering that some documents can be significantly long, larger models could take up to several minutes to output their predictions. Hence, from the available pre-trained models in Portuguese, we avoided going much further than the 335M parameters from the BERTimbau Large model.

Fine Tuning Process

BERTimbau and Albertina provide a robust foundation as they are pre-trained on large Portuguese-language corpora. In this research, this initial pre-training is augmented for legal applications through the following fine-tuning strategy. The vocabulary used by BERTimbau and Albertina is not inherently legal-specific, which is why we perform task-specific fine-tuning on a comprehensive set of legislative and legal documents. During fine-tuning, the model learns the specific terminology, phrases, and structures that are unique to legal texts. To enhance its effectiveness in the legal domain, we use a large corpus of transposed EU legislation, Portuguese national laws, and other legal materials to ensure that the model is exposed to a wide variety of legal terms and concepts. This approach

enables the expansion of the vocabulary by ensuring that the model encounters and learns the meanings of uncommon legal words and phrases during training.

The training strategy involves two key phases: masked language modeling (MLM) and fine-tuning on administrative burden detection. In the first phase, MLM, specific words in the input sequence are masked, and the model learns to predict these masked words by leveraging the context provided by the surrounding text. This phase is crucial for adapting the model to the legal domain, as it teaches the model to understand the meaning and usage of legal terminology. The second phase involves fine-tuning the model specifically for detecting administrative burdens, where it learns to identify the portions of text that impose compliance requirements, using annotated legislative texts as the ground truth (see Section 3.3).

Throughout the training process, we carefully monitor performance metrics such as loss, accuracy, and F1-score to ensure that the model is learning the task effectively without overfitting. The combination of masked language modeling and task-specific fine-tuning on legal corpora allows the model to expand its vocabulary and improve its performance in understanding legal texts.

Contextual embeddings provided by BERTimbau and Albertina play a crucial role in distinguishing terms that may appear similar at first glance but have distinct meanings in specific legal contexts, thus addressing the potential for bias in embeddings due to the use of similar features for distinct legal terms. Unlike static word embeddings, contextual embeddings allow the model to generate different representations of the same word depending on its surrounding context. For instance, in a legal sentence where the word "robbery" is used in a context involving physical force, the embedding generated by BERTimbau will reflect this specific meaning. Similarly, when the word "theft" is used in a context emphasizing unlawful appropriation without force, the embedding will capture this distinction. This context-sensitive approach is key to ensuring that the model can differentiate between terms like "robbery" and "theft" based on their legal usage.

Lastly, we employ a post-processing validation step where legal experts review the model's predictions, particularly for terms that may be prone to confusion due to their similarity. This human-AI feedback loop ensures that any misclassifications are corrected and that the model improves over time by learning from expert feedback. By continuously refining the model based on these corrections, we minimize the likelihood of biased or incorrect interpretations of legal terms in the final outputs.

3.2. Other Models

The previous section describes the BERT-based models used in this research. However, it is important to highlight that different architectures and machine learning techniques have been considered in the inception of this investigation. Our justification for selecting BERTimbau and Albertina over alternatives is as follows: encoder-based models such as BERT and Albertina are particularly well-suited for text classification tasks, as they efficiently process entire input sequences simultaneously. This makes them highly effective in detecting administrative burdens within legal texts, where contextual understanding is crucial. On the other hand, decoder-based models (e.g., GPT-3, GPT-4) [33] are optimized for text generation rather than classification. While powerful for tasks such as summarization and conversational AI, their autoregressive nature makes them computationally expensive and less efficient for classification tasks such as burden detection [34]. Additionally, sequence-to-sequence models (e.g., T5 [35]) introduce unnecessary computational overhead when applied to classification. While T5 is highly effective for translation and text generation, its dual encoding-decoding mechanism is not essential for the binary classification task at hand. Thus, given our focus on computational efficiency and scalability

for real-world applications, BERT-based models provide a more optimal trade-off between accuracy and inference speed.

Several legal-specific NLP models have been designed to handle domain-specific text, but they present key limitations. Rule-based systems, like LexNLP [36], rely on handcrafted rules and keyword matching, which limits their ability to generalize across diverse legislative formulations. In particular, they lack deep contextual understanding, leading to lower recall in detecting nuanced obligations. Pre-trained legal transformers (e.g., CaseLawBERT [37], PoLBERT [38]), while effective for common law jurisdictions, are not optimized for Portuguese legislative texts and would require extensive retraining on domain-specific corpora. In this sense, BERTimbau and Albertina were selected because they were pre-trained on extensive Portuguese corpora, capturing linguistic nuances essential for processing legislative documents in Portugal. An analysis of the performance of the different architectures is presented in Section 4.

3.3. Data

Unfortunately, there were no previously explored or processed data to be re-used for training the AI4IA system—i.e., no annotated directives for training machine learning models. Therefore, new data were acquired, and the sources for such were two-fold: (1) For the goal of building models capable of analyzing transposed legislation, documents were extracted from the Portuguese Official Journal website. (2) On the other hand, to teach models how obligations appear within EU directives that ought to be transposed, the source documents were extracted from the EUR-Lex website. Both sources are established as governmental services with universal and free access, including the possibility of printing, archiving, searching, and unrestricted access to legislative content. Given the focus on transposed EU legislation, the dataset includes both national laws and European directives that impose administrative burdens on businesses and citizens. To ensure the representativeness of the dataset, the following principles guided document selection: (1) directives that have been transposed into Portuguese law were prioritized to capture administrative burdens resulting from regulatory compliance; (2) to cover different legislative categories, the dataset includes laws, decree-laws, and ordinances to reflect different levels of legislative authority; (3) the corpus was constructed to maintain a proportional balance between Portuguese national legislation and EU directives, ensuring a diverse regulatory perspective. All the documents contain Portuguese text. To ensure the quality and reliability of the extracted text, validation procedures were applied. Specifically, the hierarchical structure of legal documents (e.g., titles, sections, and articles) was preserved to maintain context during model training. Moreover, a subset of the dataset was manually reviewed to ensure annotation correctness, preventing biases introduced during automated processing. In particular, for this task, we relied on an independent legal expert who was not involved in the labeling process (detailed in Section 3.5). Regarding their extraction, documents were acquired via programmatic procedures of collecting and processing the regulations present in the Portuguese Official Journal's website and EUR-Lex. In addition to the benefit of automating the step, the manual alternative would not be viable or, at least, scalable due to the size of the legislative corpus required to train machine learning models. Thus, custom-made programs were created for the task. The Portuguese Official Journal's website does not provide an application programming interface ("API" for short) to facilitate access to its content. Therefore, a web scraper was developed to extract the documents. The scraping application is responsible for accessing the website, searching for the desired regulation's webpage, and returning the relevant content from it. It does that through a programmatic and automated browser (Selenium), the reason being that the website's content is loaded dynamically; thus, it needs a browser to retrieve the content from the server. Otherwise, we could simply request it straight from the server. Once

collected, the program applies cleaning techniques to remove unnecessary information that accompanies the content acquired on the website (HTML tags). The process concludes with the organization of texts in a tabular structure. EUR-Lex, on the other hand, does offer an API. However, we decided to go with a similar web scraping approach, considering that the content is officially public and of free use. The motivations for such a decision were as follows: (a) the documents' webpage URLs follow a structured format that one can easily reconstruct just by having the document's name; and (b) EUR-Lex's HTML is structured in a manner that helps parse the data. It simplifies distinguishing among sections, articles, markers, and numbering, and proper text passages—a desirable aspect.

After selecting the necessary document samples from both sources and extracting the raw text from the files, these were parsed into a suitable format for training the AI models. Since the goal is to identify the specific information obligations within the document (allowing for counting and estimating them), it is preferable to work at the most granular level possible that semantically describes an individual obligation. Therefore, the legislation's texts are broken down into sentences, and each one is assigned an adequate label. In other words, each sentence-label pair represents an instance in the dataset for text classification. Section 3.3 goes into further detail regarding the labeling process conducted for the AI4IA project.

Creating Ground Truth Data

Machine learning models require annotated (or "labeled") data to learn a predictive task. Recently, multiple techniques that are aimed at mitigating the need for labeled data—be it in terms of the amount of data or the amount of time invested by the annotator—have been receiving greater attention from the scientific community. Weak supervision [39] for instance, builds training sets through labeling functions—e.g., distilled experts' annotation logic that works as computer programs (functions)— which label an entire dataset instantly (although with less precision than a manual process). Another technique is active learning [40], which minimizes the amount of manual labeling needed to achieve satisfactory results by choosing samples that lie close to the model's decision boundary and asking the human annotator to label them correctly. Considering the lack of formal rules to determine the minimum number of samples to train an AI model, we envision this step as a continuous process where we assess the performance of the model whenever a new batch of data is available. Once we reach a satisfactory performance, the manual labeling process will stop. However, none of these techniques completely eliminates the need for (at least, some) "expert approved" labeled data—a high-quality supervised sample that shall serve as ground truth for the AI models' development. Therefore, this section discusses the protocol followed to create a ground truth database for the project.

3.4. Sample Selection

The process begins by determining which documents to annotate. The labeled data should reflect as much as possible the regulations to be analyzed by the models. Since the goal is to attend stakeholders in multiple areas of public governance, it would be hard to foresee the documents' metadata, such as legislation type, impacted economic sectors, and issuer. Therefore, it was established that the investigation should be designed initially as neutral towards regulations' characteristics—i.e., documents should be randomly selected from both sources described earlier. Thus, an initial list of directives and transposed legislations was selected and annotated by two legal professionals that deeply understand the regulatory impact assessment process. Throughout the research, the model's predictions on out-of-sample data—i.e., the dataset withheld during training—were repurposed as

input for error analysis. From that, additional data were strategically labeled to supplement the cases in which the model performed poorly, and the cycle repeated until a desirable result was achieved.

3.5. Labeling

The actual data labeling procedure was conducted through a web-based application where both the domain experts and the data science team could work collaboratively. All the selected documents' text passages were combined into one dataset and loaded into the application. Figure 2 shows the annotation interface for a specific text passage.

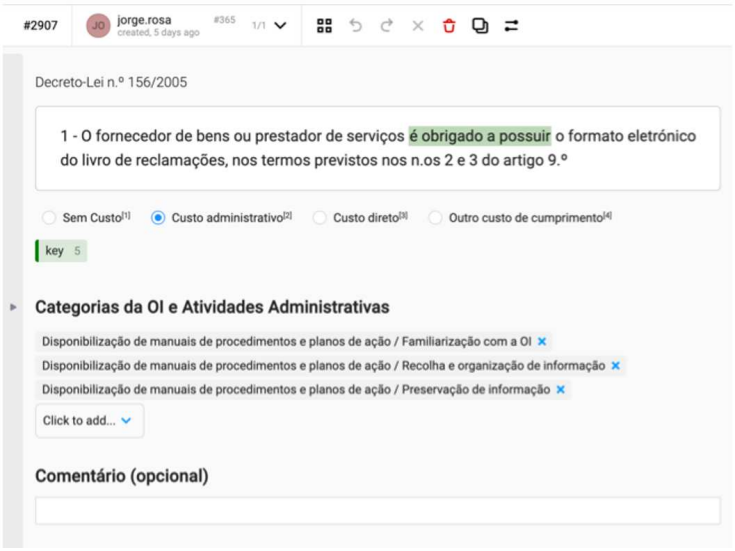

Figure 2. Text passage annotation interface of the web-based data labeling application being used for the AI4IA system.

Three inputs are requested from the expert (described below) along with an optional "comments" input:

- Cost type. For each passage, the domain expert must classify if the text indicates the existence of an administrative cost or not.
- Key phrase (only for passages marked as an administrative cost). The expert should also indicate the specific terms (or key phrase) that served as a "trigger" communicating the existence of an information obligation (e.g., "must deliver documentation" or "requires informing customers").
- Categories (only for passages marked as an administrative cost). Finally, the expert must indicate the type of IOs identified in the text.

Along with the annotations at the text passage level, experts are also expected to indicate document-level information—namely, the economic sectors the regulation impacts and other legislations it is related to. The latter refers to either the original directive from which the national legislation was transposed (or vice versa) or other previously published legislations that it alters/revokes.

Parallel to the domain experts' activities, the data science team frequently pulled the labeled data to leverage machine learning models that could help guide the process. For instance, confident learning [41] techniques were applied to highlight possible labeling errors made by the experts (usually, hard-to-interpret passages). The models could also select the most relevant passages to be labeled next by the experts—i.e., an active learning setting [42].

After 41 documents were annotated (which demanded an extensive amount of time), the combined count of text passages with administrative burdens reached only

322 examples, whilst the documents' combined corpus amounted to 19,465 text passages. In scenarios with such a significant imbalance between the target classes—on average, only 1.68% of documents' passages contained burdens—models might exhibit bias towards the majority class and ignore the minority class altogether [43]. With such a challenge at hand, techniques were explored for increasing the data's quality and reducing the need for manually labeling data. For this purpose, we built the noise reducer module. This preprocessing component of the proof-of-concept pipeline focuses on removing passages from the input corpus that were previously known to not contain administrative burdens. It follows a series of rules defined together with impact assessment experts, leveraging their prior knowledge of laws' structure. It is designed to work as an automated, algorithmic technique (e.g., instead of manual cleaning of the data), serving as a viable option to be included in the proof-of-concept pipeline when deployed to production. The algorithm's functioning is summarized as follows:

- Automatically identifies—and labels as not having obligations—titles, subtitles, numbers-only and symbols-only passages, and other content-poor text snippets.
- Documents' introductions are also disregarded, as administrative burdens are only stated within the regulation's articles.
- In the case of regulation altering an existing legal document, the module can identify only the new passages and ignore outdated ones.
- Finally, and most significantly, it extracts entire articles that indicate the absence of administrative burdens through their underlying semantic context—e.g., articles referring to the law's general logistics.

As a means of comparison, during the experimental evaluation, the noise reduction module was able to alleviate the target imbalance, increasing the average percentage of passages with administrative burdens within documents from 1.68% to 5.89%.

Finally, along with the programmatic labeling of irrelevant passages from the noise reducer, a simple normalization step was also included to standardize the inputs' format. Specifically, all markers and numbering prefixes were removed, and whitespaces were striped. Techniques such as stemming and removing stop words were avoided due to the rationale that models that work with contextual embeddings (such as transformer models) function better when the semantic context of the passage is preserved.

4. Experimental Phase

This section discusses the experimental setup for developing the main capability of the AI4IA system. Specifically, it describes the steps we took to train the NLP "filter" model for identifying the presence of administrative burdens. Figure 3 provides an overview of the process of fine-tuning the model.

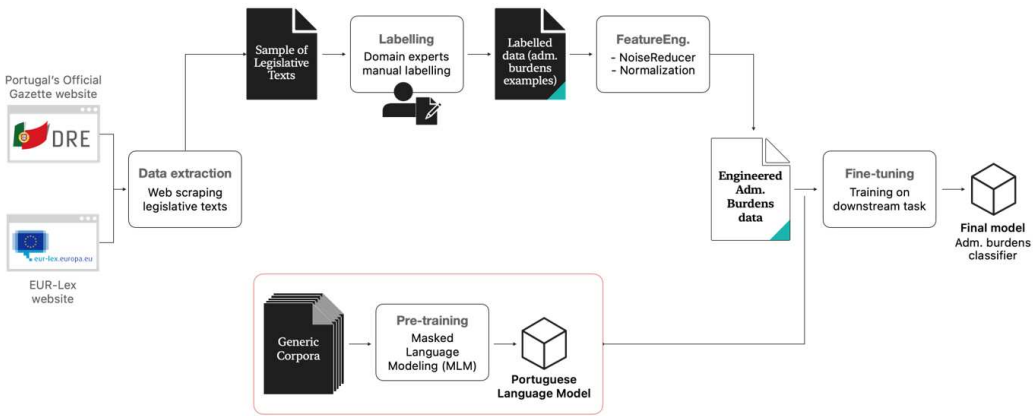

Figure 3. Systemic view of the model's development process.

In the experimental phase, we first compared the three pre-trained models with their default hyperparameter values. The goal here was to avoid a massive hyperparameter search that would result from attempting to tune values for all three pre-trained models. Afterwards, once the best model was identified at this first step, a hyperparameter search was conducted for it, followed by a final training with the best hyperparameter values. For the evaluations on unseen data, we split the data into two sets (development −80% and test −20%) using a stratified split—i.e., the target variable's distribution is maintained the same among all sets. For the hyperparameter tuning, we also conducted a stratified split to sample a validation set from the development set (i.e., the development set's 80% was further split into 60% for training and 20% for validation). The process was repeated 10 times to ensure statistical robustness. No metadata influenced the splits, such as the document's published date, for instance. That is because the envisioned usage of the proof of concept is not only to analyze newly published legislation but also the existing regulatory stock. Therefore, a uniform weight was given to the date of publication. The same rationale was applied for document type (i.e., law, decree-law, and ordinance) and all other metadata, for that matter. As mentioned above, to approximate optimal values for the algorithm's hyperparameters, a simple grid search was conducted. Specifically, as recommended in [44], epochs, batch size, and learning rate values were tuned on the validation set. For each hyperparameter, the values considered were as follows:

- Number of epochs: 2, 3, 4;
- Batch size: 16, 32;
- Learning rate: 5×10^{-5}, 3×10^{-5}, 2×10^{-5}.

These values were chosen based on empirical best practices in fine-tuning transformer-based models, particularly in low-resource classification tasks where class imbalance can be significant. Once the optimal values considered within the search were identified, the final model was built using such hyperparameter values on all the available development data—i.e., the training set and the validation set combined. This model was then assessed on the unseen data. Regarding metrics, Precision, Recall, and F1-score were used to assess the best model performance.

Given the extreme class imbalance, we adopted several strategies to ensure effective learning: First, as mentioned in previous sections, the noise reduction module pre-filtered non-informative passages, increasing the burden-containing text proportion to 5.89%. Concerning the loss function, we applied a class-weighted cross-entropy loss to penalize misclassification of the minority class more heavily. Finally, we employed data augmentation by oversampling burden-containing passages to reinforce minority class representation during training. The combination of these strategies was fundamental to overcome the extreme class imbalance that characterizes the problem.

To mitigate overfitting, we used an early stopping callback that stopped the training process if validation loss did not improve for two consecutive epochs, thus preventing unnecessary computation and avoiding overfitting.

4.1. Results Analysis

In this section, we assess the results achieved during modeling. The system aims to surpass a 0.50 Precision and 0.5 Recall for the binary task of classifying burdens. Along with such a threshold representing a performance better than a random classifier, in the case of identifying administrative burdens, that goal becomes challenging due to the extreme imbalance in the dataset. Therefore, achieving such a result validates the hypothesis of an artificial intelligence model being capable of learning the semantic context representing an information obligation. We start by reviewing the initial comparison between the pre-trained models. As one can see in Table 1, BERTimbau Base consistently outperformed

other models, offering a good balance between classification performance and computational efficiency. T5-small and GPT-2, though effective for generative tasks, showed lower performance for classification. PoLBERT, though legal domain-tuned, underperformed due to its non-Portuguese training data. These results further substantiate our choice of BERTimbau and Albertina as the most suitable models for administrative burden detection in Portuguese legal texts.

To assess whether the performance difference between BERTimbau Base and Albertina is statistically significant, we performed a paired t-test on their F1-scores across the 10 validation folds. The results yielded a p-value of 4.35×10^{-9}, indicating that BERTimbau Base significantly outperforms Albertina ($p < 0.001$). This confirms that the observed performance improvement is not due to chance and reinforces the selection of BERTimbau Base as the optimal architecture for the AI4IA system. Despite its excellent performance, the BERTimbau Base's hyperparameters were optimized. From the hyperparameter tuning results, the best model was produced with three epochs, a batch size of 32 samples, and a learning rate of 5×10^{-5} for the AdamW optimizer—i.e., the Adam algorithm with a weight decay fix [45]. These were the values considered for building the final model, which leveraged all the available development data.

Table 1. Validation set metrics for different pre-trained architectures. Confidence intervals are based on observed performance variance across the 10 folds.

Model	# Par	Language Pre-Training	Precision (±)	Recall (±)	F1-Score (±)
BERTimbau Base	110 M	Brazilian Portuguese	0.63 ± 0.02	0.77 ± 0.01	0.69 ± 0.01
BERTimbau Large	335 M	Brazilian Portuguese	0.49 ± 0.03	0.61 ± 0.02	0.54 ± 0.02
Albertina 100M PT	100 M	European Portuguese	0.57 ± 0.02	0.66 ± 0.02	0.61 ± 0.01
T5-small PT	60 M	European Portuguese	0.51 ± 0.03	0.60 ± 0.03	0.55 ± 0.02
GPT-2 PT	124 M	European Portuguese	0.50 ± 0.02	0.57 ± 0.03	0.53 ± 0.02
PoLBERT	336 M	English (legislative documents)	0.38 ± 0.03	0.42 ± 0.04	0.40 ± 0.03
BERTimbau Base (fine-tuned)	110 M	Brazilian Portuguese	0.65 ± 0.02	0.79 ± 0.01	0.71 ± 0.01

To assess the generalization ability of BERTimbau and Albertina, we assess their perfomance on a test set consisting of legislative documents not used in the training/vlaidation phase. As shown in Table 2, BERTimbau Base achieved the best result across the pre-trained models, with a Recall of 0.715, Precision of 0.62, and F1-score of 0.664. This result already outperforms the target threshold for the system.

Table 2. Performance of different models in detecting the presence of administrative burdens. Results on unseen data. "F1", "P", and "R" specify F1-scores, Precision scores, and Recall scores, respectively.

Model	F1	P	R
BERTimbau Base [initial evaluation]	0.664	0.62	0.715
BERTimbau Large [initial evaluation]	0.519	0.470	0.575
Albertina 100 M PT [initial evaluation]	0.582	0.552	0.615
BERTimbau Base' [final evaluation]	0.678	0.611	0.762

The superior performance of the BERTimbau Base model can be attributed to several properties of transformer architectures: the self-attention mechanism allows the model to focus on relevant parts of the legislative text dynamically, capturing long-range dependencies and contextual relationships that are crucial for understanding complex legal language. Moreover, as previously mentioned, the model was pre-trained on the extensive brWaC

dataset, which includes 2.7 billion tokens. This extensive pre-training phase enables the model to acquire a deep understanding of the Portuguese language, significantly boosting its performance on downstream tasks. The fine-tuned model shows a Recall of 0.762, Precision of 0.611, and F1-score of 0.678.

In other words, the system is currently able to identify, on average, 76.2% of the administrative burdens present within a legal document (i.e., Recall score). On the other hand, the Precision score indicates that from all passages returned by the system, 61.1% actually contains an information obligation. Although this indicates that a fair amount of the predictions will be false positives, for the day-to-day use of the proof of concept, one could argue that it is preferable to have a higher Recall even if it means sacrificing some Precision. The reason is that it is less time-consuming to review whether the identified passages have an administrative burden than look for those missed by the system within the entire text.

In addition to transformer-based models, we evaluated the performance of traditional machine learning approaches (e.g., support vector machines (SVMs), decision trees, and random forest) on the administrative burden detection task. These models have been widely used in NLP but present significant generalization challenges when applied to complex legal texts. Specifically, classical ML methods rely on handcrafted features, such as TF-IDF and word n-grams [46], which are insufficient for capturing the semantic structures of legislative language. Moreover, by treating words as independent features, they ignore sentence structure and context, making it difficult to detect implicit administrative burdens.

In our experiments, traditional models such as SVM and random forest achieved F1-scores of approximately 0.45, significantly lower than the 0.678 F1-score of BERTimbau Base. Moreover, recall scores for classical models remained below 0.50, indicating their limited ability to identify burdens comprehensively.

Thus, while traditional machine learning approaches may work for simpler NLP tasks, they are insufficient for administrative burden detection in legislative texts. BERTimbau and Albertina provide a domain-adapted solution with superior performance in understanding and classifying legislative obligations, making them the optimal choice in the context of the AI4IA project.

To assess the combined performance of the human expert supported by the AI4IA system, rigorous experiments were conducted with the collaboration of three independent experts. Initially, we selected ten documents representing national legislation derived from transposed EU directives, each previously annotated with ground truth examples of administrative burdens. The first step involved training the three experts on the AI4IA system to ensure a consistent understanding of its functionalities and the specific criteria for identifying administrative burdens. Each expert was tasked with analyzing the same set of documents through the system's graphical user interface. They were provided with the preliminary report generated by the AI4IA system as a foundational reference. This report included identified administrative burdens, which the experts then reviewed critically. During their analyses, the experts were encouraged to provide comprehensive feedback on the identified burdens (this included both removing any inaccurately flagged information obligations and adding any additional obligations they deemed relevant). Following their independent evaluations, the feedback from all three experts was synthesized to create a consolidated final list of identified and validated administrative burdens. This collaborative approach not only mitigated potential bias from a single annotator but also enriched the analysis through diverse expert perspectives.

To assess the effectiveness of the AI4IA system in supporting the expert reviews, we compared the final consolidated list against the document's ground truth. We calculated the average number of administrative burdens identified by each expert, allowing us to

determine the system's Recall. The results of these assessments, broken down by document, are presented in Table 3. This evaluation process ensured that multiple expert viewpoints informed the identification of administrative burdens, thus reinforcing the validity of our findings.

Table 3. Experiments conducted with an impact assessment consultant analyzing transposed Portuguese legislation with the support of the AI-based system.

Document	Ground Truth's Burdens Count	AI + Expert Burdens Count	Expert's Analysis Elapsed Time	Observation
Decreto-Lei n.º 87/2018	5	5	00:10:20	-
Decreto-Lei n.º 80/2017	16	17	00:18:05	The expert identified one additional burden not considered in the ground truth
Decreto-Lei n.º 17/2018	10	8	00:08:40	-
Decreto-Lei n.º 78/2018	2	5	00:05:38	The expert identified three additional burdens not considered in the ground truth
Lei n.º 32/2019	2	0	00:10:54	The expert adjusted (removed) the two ambiguous burdens considered in the ground truth
Decreto-Lei nº 109-G/2021	23	20	00:17:42	-
Decreto-Lei nº 84/2021	7	7	00:15:50	-
Decreto-Lei nº 73/2020	8	6	00:16:21	-
Decreto-Lei nº 225/2006	5	5	00:09:43	-
Portaria n.º 1320/2008	7	7	00:12:48	-

AI4IA has certain failure cases, which we have analyzed through manual review of misclassified examples. In particular, the model may incorrectly classify information obligations in two cases: (1) if they contain compliance-related terminology without actually imposing an obligation; and (2) if they contain unusual legal formulations, such as highly specialized sectoral regulations, may be misclassified due to limited training data. Table 4 presents passages that were misclassified by the system, along with proposed strategies to mitigate these errors.

Table 4. Examples of passages missclassified by the AI4IA system. Examples adapted from Portuguese text.

Example Type	Example Passage	Ground Truth	Model Prediction	Potential Cause	Recommended Mitigation
False Positive	"The competent authority shall record the date of receipt".	No burden	Burden	Procedural language mimics compliance phrasing	Expand training data with procedural non-burdens; enhance negative examples
False Positive	"All stakeholders were consulted prior to drafting the law".	No burden	Burden	Historic procedural event misclassified as current requirement	Extend non-burden examples from legislative history and process narratives
False Positive	"The Ministry will publish the findings annually".	No burden	Burden	Use of authoritative verb ("will publish") mistaken for obligation	Train on reporting statements to distinguish declarative vs. directive tones
False Negative	"Entities shall, upon request, provide documentation to oversight bodies".	Burden	No burden	Obligation masked by conditional clause ("upon request")	Incorporate more conditional phrasing in training samples
False Negative	"Registrations must be completed no later than 30 days after issuance of license".	Burden	No burden	Time-constrained requirements embedded in passive phrasing	Annotate more passive-form deadlines and temporal requirements
False Negative	"Institutions are expected to comply with data-sharing protocols".	Burden	No burden	Soft obligation phrased using modal expression ("are expected to")	Expand annotations with soft-mandate patterns like "expected to", "recommended to"

However, the combined effort of the AI4IA system and human experts has proven highly effective in detecting the vast majority of existing administrative burdens. By significantly accelerating the assessment process, the system reduces the time required for impact evaluation from several hours (or even days) to just a few minutes. This rapid "turnaround" not only enhances the efficiency of human experts but also allows for more timely and informed decision-making. Moreover, the proposed system's usefulness extends

beyond speed; it enhances accuracy and consistency in the assessment process, mitigates subjective biases, and frees up human experts to focus on more complex and nuanced aspects of legislative analysis. Consequently, the AI4IA system represents a transformative tool in the legislative impact assessment domain, offering substantial improvements in productivity, reliability, and transparency. To provide a complete picture of the system's deployment feasibility, we describe the hardware and resource requirements of the AI4IA system. The analysis of legislative documents is currently executed on an instance with eight virtual CPU cores (vCPUs) and 8 GB of RAM, which has proven sufficient for smooth operation (i.e., lengthy documents are analyzed in approximately five minutes). Any specifications above this configuration are only necessary for added throughput or safety margin, not for baseline functionality. From a storage perspective, hosting 20 legislative documents requires approximately 6.31 MB in the database. This demonstrates that AI4IA is suitable for scalable deployment. Overall, the system can be deployed using standard institutional or cloud-based infrastructure, without requiring specialized GPU hardware or high-performance computing environments.

4.2. Reducing Administrative Burdens and Time Savings with AI4IA

The reduction of administrative burdens is a critical concern for governments and regulatory bodies worldwide, particularly in the context of legislative impact assessments. These burdens often manifest as compliance costs for businesses, organizations, and individuals, which result from obligations such as maintaining records, filling out forms, and submitting reports to government authorities.

The AI4IA system offers a powerful solution to this problem by automating the detection of administrative burdens in legislative texts, significantly reducing the manual effort required from legal experts. Traditionally, the task of reviewing legislation for potential administrative burdens has been a time-consuming and labor-intensive process, involving the careful examination of legal texts to pinpoint provisions that may impose compliance obligations. This manual approach is not only resource-intensive but also prone to variability, as different reviewers may interpret the same text differently.

One of the most significant benefits of the AI4IA system is the reduction in time required for legal experts to complete legislative impact assessments. Empirical results from the system's evaluation show that it can significantly reduce the time needed to review legislative documents for administrative burdens. These time savings are critical in environments where legal experts are often tasked with reviewing large volumes of legislative text under tight deadlines. By automating the initial detection of administrative burdens, the AI4IA system allows legal experts to focus on higher-level tasks, such as interpreting the broader implications of the legislation, drafting amendments, or advising policymakers on strategic regulatory decisions. The system thus not only improves efficiency but also enhances the quality of the legislative process by freeing up expert resources for more complex tasks.

The impact of time savings extends beyond the individual legal experts or government agencies conducting the reviews. By streamlining the legislative impact assessment process, the AI4IA system can accelerate the entire legislative cycle, allowing new or amended legislation to be drafted, reviewed, and implemented more quickly. This speed is particularly important in contexts where regulations must keep pace with fast-evolving industries, such as technology, healthcare, and environmental protection. In these sectors, delays in the legislative process can lead to regulatory gaps, where outdated or incomplete laws fail to address emerging challenges. The AI4IA system helps mitigate these risks by ensuring that the legislative process is both efficient and responsive to changing societal needs.

Moreover, the automation provided by the AI4IA system enhances the consistency and objectivity of the legislative impact assessment process. Manual reviews are inherently subjective, as different reviewers may prioritize certain types of burdens over others or may interpret the text differently based on their experience and expertise. The AI4IA system provides a uniform approach to identifying administrative burdens, applying the same criteria and algorithms across all texts. This uniformity reduces the likelihood of errors or omissions and ensures that all burdens are treated with equal scrutiny, contributing to a more transparent and accountable legislative process.

In summary, the AI4IA system offers significant advantages in terms of reducing administrative burdens and saving time in the legislative impact assessment process. By automating the detection of burdensome clauses in legal texts, the system alleviates the workload on administrative personnel and legal experts, allowing them to focus on more strategic and high-level tasks. The resulting time savings not only benefit individual reviewers but also contribute to a faster, more efficient legislative cycle, improving the responsiveness and effectiveness of regulatory frameworks. As governments and regulatory bodies continue to seek ways to streamline their processes, the AI4IA system provides a solution for reducing administrative burdens and improving legislative outcomes.

5. Conclusions

The main objective of this research was to develop an AI-based system to automate the legislative impact assessment (LIA) process by specifically targeting the detection of administrative burdens from transposed EU legislation. To achieve this, we employed a methodology centered around a BERT-based model. This model, initially trained on common Portuguese language data, was fine-tuned for legislative documents to ensure high accuracy and relevance in detecting administrative burdens.

The implementation of this system has significant implications for legislative impact assessment experts. By automating the detection of administrative burdens, the system allows experts to save substantial time, which can be redirected towards other critical tasks. Furthermore, it enables the production of more timely and efficient LIA reports, enhancing the overall effectiveness of the legislative assessment process.

While AI4IA demonstrates strong performance, several challenges remain. First, legal language is complex, with obligations often expressed implicitly. Thus, while contextual embeddings improve understanding, certain legislative nuances may still be misclassified. The model's robustness across different formulations requires ongoing refinement with expert feedback. Second, the system is fine-tuned for Portuguese legislation, and its generalization to other EU jurisdictions requires additional training on multilingual and jurisdiction-specific legal corpora. For this reason, future work will explore transfer learning techniques to adapt the model to new legal frameworks. Finally, while AI4IA achieves high recall, false positives remain a concern, requiring expert intervention. Conversely, some administrative burdens may remain undetected, particularly those embedded in indirect legal references.

To address these challenges, future research will focus on expanding the training corpus to include legal texts from additional jurisdictions, improving cross-border applicability. Moreover, considering that legislative texts evolve (thus necessitating frequent retraining to maintain model accuracy), future developments should incorporate active learning to enable the system to dynamically learn from new legislative texts with minimal manual intervention.

To expand the capabilities of the proposed AI4IA system, several research ideas can be explored. One potential direction is to extend the impact assessment to cover social and environmental aspects, providing a more holistic evaluation of proposed legislation. Addi-

tionally, incorporating dynamic impact assessment mechanisms would allow the system to adapt to changing conditions and continuously provide relevant insights. These enhancements would further increase the value of the AI-based system, making it a fundamental tool for performing a comprehensive legislative impact assessment exercise.

Author Contributions: Conceptualization, V.C., M.C., and P.C.; investigation, V.C.; methodology, V.C.; software, V.C.; supervision, M.C. and P.C.; validation, V.C., M.C., and P.C.; writing—original draft, V.C.; writing—review and editing, V.C., M.C., and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project-UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020)-Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS), and the project 2024.07277.IACDC (Lexa). This work was supported the European Union through the project TSI-2022-AI4IA-EU-IB-22PT09-Artificial intelligence for better regulation. However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code of the AI4IA system cannot be shared due to existing restrictions imposed by the beneficiary of the TSI-2022-AI4IA-EU-IB project. For more information on the system, the interested reader can contact Victor Costa (vcosta@novaims.unl.pt).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kirkpatrick, C.H.; Parker, D. *Regulatory Impact Assessment: Towards Better Regulation?* Edward Elgar Publishing: London, UK, 2007.
2. Dunlop, C.A.; Maggetti, M.; Radaelli, C.M.; Russel, D. The many uses of regulatory impact assessment: A meta-analysis of EU and UK cases. *Regul. Gov.* **2012**, *6*, 23–45.
3. Alemanno, A. How much better is better regulation? Assessing the impact of the better regulation package on the European Union—A research agenda. *Eur. J. Risk Regul.* **2015**, *6*, 344–356. [CrossRef]
4. Simonelli, F.; Iacob, N. Can we better the European Union better regulation agenda? *Eur. J. Risk Regul.* **2021**, *12*, 849–860.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
6. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132.
7. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
8. Zhang, H.; Shafiq, M.O. Survey of transformers and towards ensemble learning using transformers for natural language processing. *J. Big Data* **2024**, *11*, 25.
9. Krasadakis, P.; Sakkopoulos, E.; Verykios, V.S. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics* **2024**, *13*, 648. [CrossRef]
10. European Comission Action Programme for Reducing Administrative Burdens in the EU Final Report. Accompanying the Document Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. EU Regulatory Fitness. COMMISSION STAFF WORKING DOCUMENT. SWD (2012). 2012. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52012SC0423 (accessed on 9 January 2025).
11. Hradec, J.; Ostlaender, N.; Macmillan, C.; Acs, S.; Listorti, G.; Tomas, R.; Arnes Novau, X. *Semantic Text Analysis Tool: SeTA*; Publications Office of the European Union: Luxembourg, 2019.
12. Levy, O.; Goldberg, Y. Neural word embedding as implicit matrix factorization. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
13. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
14. Ostlaender, N.; Acs, S.; Listorti, G.; Hardy, M.; Ghirimoldi, G.; Hradec, J.; Smits, P. *Modelling Inventory and Knowledge Management System of the European Commission (MIDAS)*; Publications Office of the European Union: Luxembourg, 2019.

15. Acs, S.; Ostlaender, N.; Listorti, G.; Hradec, J.; Hardy, M.; Smits, P.; Hordijk, L. *Modelling for EU Policy Support: Impact Assessments*; Technical Report; Publications Office of the European Union: Luxembourg, 2019.
16. Misuraca, G.; Van Noordt, C. *AI Watch-Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU*; JRC Working Papers; Publications Office of the European Union: Luxembourg, 2020; ISBN 978-92-76-19540-5.
17. Romic, D.; Halak, Z.V. Regulatory Impact Assessment in the Republic of Croatia-Situation and Perspective. *Chall. Knowl. Soc.* **2014**, 880–891.
18. Zeitz, D. *Better Regulation in Germany as Quality Assurance System: Recent Development and Current Challenges*; Netherlands Administrative Law Library (Nall): Utrecht, The Netherlands, 2016.
19. Karpen, U. Regulatory impact assessment: Current situation and prospects in the German Parliament. *Amic. Curiae* **2015**, *101*, 14.
20. Kosach, I.; Shaposhnykov, K.; Chub, A.; Yakushko, I.; Kotelevets, D.; Lozychenko, O. Regulatory policy in the context of effective public governance: Evidence of Eastern European Countries. *Cuest. Políticas* **2022**, *40*, 1–20.
21. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624.
22. Wang, H.; Li, J.; Wu, H.; Hovy, E.; Sun, Y. Pre-trained language models and their applications. *Engineering* **2023**, *25*, 51–65. [CrossRef]
23. Lipton, Z.C.; Berkowitz, J.; Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv* **2015**, arXiv:1506.00019.
24. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019.
25. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62.
26. Rodrigues, J.; Gomes, L.; Silva, J.; Branco, A.; Santos, R.; Cardoso, H.L.; Osório, T. Advancing neural encoding of portuguese with transformer albertina pt. In Proceedings of the EPIA Conference on Artificial Intelligence, Faial Island, Portugal, 5–8 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 441–453.
27. Souza, F.; Nogueira, R.; Lotufo, R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In Proceedings of the Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, 20–23 October 2020; Proceedings, Part I 9; Springer: Berlin/Heidelberg, Germany, 2020; pp. 403–417.
28. Lan, Z. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
29. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
30. Alyafeai, Z.; AlShaibani, M.S.; Ahmad, I. A survey on transfer learning in natural language processing. *arXiv* **2020**, arXiv:2007.04239.
31. Koroteev, M.V. BERT: A review of applications in natural language processing and understanding. *arXiv* **2021**, arXiv:2103.11943.
32. Santos, R.; Rodrigues, J.; Gomes, L.; Silva, J.; Branco, A.; Cardoso, H.L.; Osório, T.F.; Leite, B. Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family. *arXiv* **2024**, arXiv:2403.01897.
33. Kalyan, K.S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2024**, *6*, 100048.
34. Shaheen, Z.; Wohlgenannt, G.; Filtz, E. Large scale legal text classification using transformer models. *arXiv* **2020**, arXiv:2010.12871.
35. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
36. Bommarito II, M.J.; Katz, D.M.; Detterman, E.M. LexNLP: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on Big Data Law*; Edward Elgar Publishing: London, UK, 2021; pp. 216–227.
37. Zheng, L.; Guha, N.; Anderson, B.R.; Henderson, P.; Ho, D.E. Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, São Paulo, Brazil, 21–25 June 2021.
38. Henderson, P.; Krass, M.; Zheng, L.; Guha, N.; Manning, C.D.; Jurafsky, D.; Ho, D. Pile of law: Learning responsible data filtering from the law and a 256 gb open-source legal dataset. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 29217–29234.
39. Ratner, A.; Hancock, B.; Dunnmon, J.; Sala, F.; Pandey, S.; Ré, C. Training complex models with multi-task weak supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4763–4771.
40. Settles, B. From theories to queries: Active learning in practice. In Proceedings of the Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010, Sardinia, Italy, 16 May 2011; JMLR Workshop and Conference Proceedings; pp. 1–18.
41. Northcutt, C.; Jiang, L.; Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* **2021**, *70*, 1373–1411.
42. Ren, P.; Xiao, Y.; Chang, X.; Huang, P.Y.; Li, Z.; Gupta, B.B.; Chen, X.; Wang, X. A survey of deep active learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40.

43. Shakeel, F.; Sabhitha, A.S.; Sharma, S. Exploratory review on class imbalance problem: An overview. In Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
44. Feurer, M.; Hutter, F. *Hyperparameter Optimization*; Automated Machine Learning Methods Systems Challenges; Springer International Publishing: Cham, Switzerland, 2019; pp. 3–33.
45. Loshchilov, I.; Hutter, F. Fixing weight decay regularization in adam. *arXiv* **2017**, arXiv:1711.05101.
46. Shirakawa, M.; Hara, T.; Nishio, S. Idf for word n-grams. *ACM Trans. Inf. Syst. (TOIS)* **2017**, *36*, 1–38. [CrossRef]