Automatic Classification of Access Levels in Documents from the Brazilian Electronic Information System

Ana Clara B. Borges[1], Mayara C. Marinho[2], Rodrigo de Freitas Nogueira[3], Jacir L. Bordim[2], Vinicius R. P. Borges[2]

[1] Faculdade de Ciências e Tecnologias em Engenharia – Universidade de Brasília (UnB)
Brasília – DF – Brazil

[2]Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Brasília – DF – Brazil

[3]Arquivo Central – Universidade de Brasília (UnB)
Brasília – DF – Brazil

{borges.anacb, mayarachewm}@gmail.com,

{rodrigofn, bordim, viniciusrpb}@unb.br

Abstract. *The Brazilian Electronic Information System (SEI) is used to manage documents and administrative processes in public institutions. Users often need to upload external documents and manually assign an access level: public, restricted, or confidential. To reduce human errors in this process, we propose an automatic classification approach. A labeled corpus was constructed using public SEI's documents and artificial samples generated from SEI's form templates filled with fictional content using ChatGPT. We trained and evaluated four classification models: SVM, LSTM, BiLSTM, and BERT. Experiments were performed by means of Stratified K-Fold cross-validation and the results showed that SVM and BiLSTM performed best, each achieving a macro F1-score of 0.96.*

1. Introduction

In 2009, the Brazilian Information System (Sistema Eletrônico de Informações - SEI) was launched as a platform for document creation and procedural management in Brazil's public administration. SEI was designed to promote the efficiency of electronic document management, and its usefulness in institutional activities was confirmed by users' perception [Camarinha et al. 2023]. Besides that, the system replaces paper-based processes, enhances communication, and promotes transparency across Brazilian society [Gottschalg-Duque 2024]. Several federal public bodies, including universities, ministries, and law enforcement agencies, have adopted SEI to modernize their workflows.

When creating a new process in SEI, users must first select a process type and provide essential metadata, such as subject, interested parties, and classification codes. Once the process is created, documents can be added in two main ways. The first is by using SEI's existing document templates, which allow users to generate standardized documents directly on the platform. These templates include institutional forms, such as requests, memorandums, and official letters, which can be filled out and digitally signed within the system. Alternatively, users can upload external documents or those created

outside SEI, such as supporting documents (e.g., certificates, flight tickets, proof of address). Each created process receives a unique identifier number, which ensures traceability and proper management throughout its entire workflow.

According to the SEI's instructional manual, an access control approach is applied to external documents uploaded to SEI. When adding an external document to a SEI's process, users must assign it one of three access levels: public, restricted, or confidential. This categorization determines the visibility and handling of the document within the system. Public documents are the only category accessible to users with a SEI's account and those without a registered account. Restricted documents are limited to users affiliated with the organizational units responsible for processing them. Confidential documents, on the other hand, require explicit authorization, ensuring that only individuals with specific permissions can access them. The handling and processing of SEI's documents, whether manual, automatic, or based on intelligent strategies, must consider that such documents may contain sensitive information and must comply with the Lei Geral de Proteção de Dados (LGPD), Brazil's General Data Protection Law. However, a lack of awareness or understanding of the legislation can lead to incorrect assessments of document sensitivity, occasionally making confidential content inadvertently accessible to the public. Conversely, some non-confidential documents that could enhance transparency remain unnecessarily restricted within the platform.

Regarding the analysis of confidential information in text documents, a significant amount of research has focused on the detection and anonymization of sensitive data [Aldeen et al. 2015, Juez-Hernandez et al. 2023], particularly in the medical field, where the presence of personal information in patient records raises significant privacy concerns [Ghann et al. 2022, Wu et al. 2021]. However, while the task of automatically categorizing external SEI's documents can be approached as a text classification problem, there is still limited research in the literature addressing document classification by privacy level [Alparslan et al. 2011, Sangaroonsilp et al. 2023b].

Text classification is a Natural Language Processing (NLP) task that aims to predict the categorical values associated with textual samples [Coelho et al. 2022]. In the context of SEI, a text classification approach could be employed to automatically assign an access level to documents, serving as a secondary verification step in the labeling process and helping reduce human errors. However, to the best of our knowledge, no publicly available corpus exists for classifying documents by access level. This step is necessary to enable the training of classification models that learn the textual patterns associated with each access level.

These gaps motivated us to propose a classification approach for automatically categorizing SEI's documents into the three aforementioned access levels. For that purpose, we curated a labeled corpus of SEI's documents inspired by those typically processed at the University of Brasília (UnB). Due to confidentiality constraints that prevented the use of real restricted and confidential documents, we generated artificial samples using an LLM and official SEI form templates as guidance. We then trained four classification models - Support Vector Machines (SVM), Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber 1997], Bidirectional LSTM [Graves 2012], and Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al. 2019], to validate the constructed corpus and evaluate their performance.

The main contribution of this paper is the approach for constructing the labeled corpus, constituted by 606 documents in Brazilian Portuguese and English associated with the three access levels. An additional contribution is the strategy for generating fictional content by filling out empty SEI's form templates of the UnB using ChatGPT, while ensuring compliance with confidentiality regulations such as Brazil's Freedom of Information Law (Lei de Acesso à Informação - LAI) and the LGPD.

The structure of this paper is as follows. Section 2 presents the related work. Section 3 describes the proposed method and its steps: data extraction and generation, corpus construction, text preprocessing, and the classification approach used. Section 4 presents the experiments and results, detailing the performance of each classification model. Finally, Section 5 summarizes the conclusions of this study and outlines directions for future work.

2. Related Work

Although most NLP research in the literature focuses on analyzing anonymization strategies in text documents [Sousa and Kern 2023, Csányi et al. 2021], this study selects works that investigate the application of NLP techniques for legal documents as well as classification approaches. The scarcity of classification approaches for assigning access levels to documents is likely due to the lack of publicly available corpora containing confidential texts. Publishing such data may violate privacy laws, disregard institutional regulations, or expose sensitive information that could compromise individuals or public interests.

Araujo et al.[De Araujo et al. 2020] applied NLP and Machine Learning techniques to Supremo Tribunal Federal (STF), Brazil's Supreme Court, documents. The authors were motivated by the challenges posed by the long average processing time of lawsuits and the large volume of cases. Their main contribution was VICTOR, a labeled corpus of legal documents from STF. The corpus supported two tasks: document type classification and theme assignment. Their method involved extracting text from PDF files and comparing different approaches for document type classification. Experiments using Conditional Random Fields (CRF) demonstrated that the sequential nature of lawsuits can be leveraged to improve classification performance.

Positioned in the context of the automated analysis of legal opinions related to consumer complaints, Coelho et al.[Coelho et al. 2022] focused on the issue of identifying the moral damage value in these legal opinions. To address this task, the authors adopted a classification approach and introduced a new model, MuDEC (Multistep Document Embedding-Based Classifier), that combines Doc2vec and Support Vector Machines (SVM). The experiments used a dataset constituted by 193 manually annotated Brazilian Portuguese legal opinions from the State Court of Rio de Janeiro labeled in six different classes. The results of their proposed method are especially relevant when compared to similar text classification approaches.

Motivated by the growing concern for data protection among governments and citizens, Sangaroonsilp et al. [Sangaroonsilp et al. 2023a] proposed an automatic method for classifying issue reports according to predefined privacy requirements. The proposed method consisted of three main steps: textual data preprocessing, feature extraction, and the training of a classification model. The classification methods were grouped into three

categories: naive baselines, traditional word embedding techniques, and deep learning approaches. The evaluation showed that Bag-of-Words, N-gram IDF, Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec are suitable techniques for classifying privacy requirements, with N-gram IDF yielding the highest classification performance.

Suvlavko et al.[Sulavko et al. 2024] proposed an ensemble of convolutional neural networks (CNNs) for detecting confidential information in text messages. A dataset in Russian was generated containing texts from 10 organizations, each labeled with one of four confidentiality levels. The ensemble includes a pre-trained E5-based model, an autoencoder to separate public and private content, and multiple CNNs for final classification. The models were trained on organizational data containing confidential documents. The proposed system achieved low error rates around 0.015% for missed confidential information and 4.3% for false detections.

The literature presents works on document classification in Brazilian Portuguese, as well as studies that explored alternative document classification strategies. However, there are no publicly available corpora with characteristics similar to SEI's documents that can be used to train classification models for predicting their access level. Inspired by previous research, this paper aims to fill this gap by introducing a bilingual corpus in English and Brazilian Portuguese, composed of public administration documents categorized by access level.

3. Proposed Method

This section presents the NLP pipeline to classify the private level of the documents from SEI. Subsection 3.1 presents the methods to obtain the public documents used. Subsection 3.2 explains the restricted and confidential documents acquisition and generation. Subsection 3.3 presents the corpus constructed. Subsection 3.4 explains the text preprocessing. Subsection 3.5 describes the classification approach and the underlying models.

The proposed method consists of the following steps: extraction of public documents from the UnB's SEI, generation of restricted and confidential documents, creation of the corpus, setup of classification models, stratified K-Fold cross validation and results analysis. The proposed method is illustrated in Figure 1.

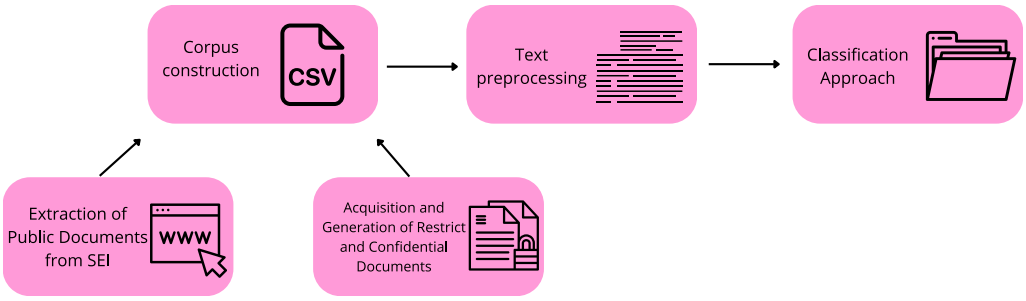

Figure 1. Proposed methodology for classifying SEI documents by access level.

3.1. Extraction of Public Documents from SEI

To collect a large number of documents published in the public section of SEI, a web crawler was developed to retrieve documents from the UnB. The crawler was imple-

mented in Python using Selenium. It accesses the public search page of the UnB's SEI, navigates to the document pages, and extracts their textual content. The extracted texts are directly labeled as public documents.

3.2. Restricted and Confidential Documents

We consulted experts on the SEI system at the UnB to gather information about the types of documents that must be categorized as restricted or confidential. Although some overlap may exist between these categories, we differentiate them as described below:

- **Restricted documents**: These are visible only to users affiliated with the organizational units involved in the SEI process. As such, they are inaccessible to the general public and to users from unrelated departments. Restricted documents typically address internal administrative matters that, while not presenting a threat to individual rights, still require limited visibility to protect institutional integrity and internal decision-making. Examples include preliminary reports, draft proposals, proof documents, invoices, certificates, budget justifications under evaluation.
- **Confidential documents**: These documents generally do not circulate between organizational units within SEI. Traditionally, they are shared individually through access credentials granted to specific users, making them inaccessible to individuals not directly involved in the process. In compliance with Brazil's Freedom of Information Law (Lei de Acesso à Informação - LAI), these documents are confidential due to their potential to compromise national or societal security. Such documents often contain sensitive personal data and consist of memorandums or official dispatches that address delicate subjects with significant potential to harm the personal, professional, or academic life of those involved.

The ethical constraints associated with using restricted and confidential access-level documents in SEI led us to generate artificial samples by leveraging the capabilities of Large Language Models (LLMs). For that purpose, we used publicly available empty templates of administrative forms from the UnB and prompted ChatGPT to populate them with fictional data. This strategy is inspired by the successful application of LLMs in the literature for generating synthetic samples for text augmentation [Long et al. 2024, Bayer et al. 2022], as well as for addressing class imbalance in datasets [Cai et al. 2023, Cloutier and Japkowicz 2023]. All formulated prompts were reviewed by a NLP expert to ensure that ChatGPT generated texts with varied writing styles and realistic content.

For the restricted access level, we randomly collected various documents and images from the Internet, including certificates, declarations, flight tickets (used as proof of travel in leave of absence cases), utility bills (water and electricity), voter compliance certificate, research proposals, invoices, letters of recommendation, and invitation letters. We used Tesseract [Smith 2007] to extract raw text from the image files and manually replaced fields that could identify a person or organization - such as names, email addresses, phone numbers, and national identification numbers (e.g., CPF) with fictional data. To preserve a sense of realism, the names of public events (e.g., conferences and workshops) were retained in the documents.

Furthermore, we collected *12* different empty form templates from the UnB, related to requests commonly submitted by students and public servants. Examples of student-related forms include requests for financial support to attend academic events,

grade review petitions, research and/or extension scholarship agreements, and reinstatement. For public servants, examples include leave of absence forms, graduate program accreditation requests, research project work plans, and travel reports. We generated between 5 and 25 documents from each type of template. We next provide a sample prompt used to fill out an empty form template that ensures variability in the generation of different versions of the filled forms:

> "You are a graduate student and you are filling out a request form for a leave of absence from a course. Follow the model below, filling in the terms that are in brackets with fictional information in Portuguese (Brazil) with fictional data.
> SOLICITAÇÃO DE TRANCAMENTO DA PÓS-GRADUAÇÃO
> 1. IDENTIFICAÇÃO DO(A) DISCENTE
> Nome completo: [Nome do discente]
> E-mail: [Email do discente]
> 2. SOLICITAÇÃO
> Período/Ano do trancamento: [Semestre/Ano do trancamento]
> Solicito: (X) o Trancamento DA(S) DISCIPLINA(S), abaixo:
> (Art. x Resolução x)
> CÓDIGO DA DISCIPLINA [Código da Disciplina]
> NOME DA DISCIPLINA [Nome da Disciplina]
> TURMA [Número da turma]
> ( ) o Trancamento GERAL de matrícula.
> (Art. x Resolução x)
> Motivo:
> ( ) Saúde do(a) discente. (documento obrigatório: comprovante(s) médico(s) e/ou psicológico(s)
> ( ) Licença maternidade - Resolução x. (documento obrigatório: certidão de nascimento)
> ( ) Outro(s). (documento obrigatório: comprovante do impedimento) *Neste caso, utilizar o formulário de exposição de motivos para especificar.
> 4. DECLARAÇÃO / ASSINATURA DO(A) DISCENTE
> [Cidade] [Dia], de [Mês], de [Ano]."

Regarding confidential documents, we followed expert recommendations to generate texts simulating reports and complaints typically submitted to institutional ombudsman channels (Ouvidoria). For this purpose, we collected complaint texts from Reclame Aqui[1], a widely used Brazilian consumer protection platform where users publicly report negative experiences with companies and public institutions. Specifically, we focused on complaints related to federal educational institutions. In addition, we formulated prompts for ChatGPT to generate complaints and reports that could be addressed by the Ouvidoria, varying the writing style from informal to formal. These included cases of moral and sexual harassment, absence from the workplace, workplace accidents, infrastructure issues, and administrative misconduct. Finally, we collected images of medical, psychological, and psychiatric reports from the Internet to obtain their blank templates, which were then filled with fictional patient data using ChatGPT.

Although we collected and generated several SEI-related documents from the Internet, the frequency distribution of documents per category was defined based on expert guidelines and is presented in Table 1.

[1] https://www.reclameaqui.com.br/

3.3. Corpus Construction

After collecting documents for the "Public", "Restricted", and "Confidential" access levels, we consolidated them into a single file to form the final corpus. Table 1 provides a detailed view of the corpus, specifying the number of documents for each access level. The proportion of confidential documents is lower compared to the other categories, both due to the difficulty in obtaining them and to reflect the real-world distribution of such documents in SEI's processes, as indicated by domain experts.

Table 1. Number of instances per category (class label).

Group	Number of documents	Approximate proportions
Public	288	47.5%
Restricted	277	45.7%
Confidential	41	6.8%
Total	606	100%

3.4. Text preprocessing

After constructing the corpus, the texts were preprocessed to prepare them for machine learning model training. The preprocessing involved the following four steps:

1. Lowercasing all text;
2. Tokenization: splitting the text into individual words (tokens);
3. Stopword removal: eliminating common but semantically insignificant words in Brazilian Portuguese and English;
4. Lemmatization: reducing words to their base form to normalize linguistic variations. This process, performed using the Python NLTK[2] library, improves text consistency and supports more accurate analysis.

3.5. Classification Approach

Four different models were developed for the classification task: a classical Support Vector Machine (SVM), two neural models based on Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), and a pre-trained BERT model.

The SVM model was employed as one of the classification algorithms due to its successful application in text classification tasks. For that purpose, input texts were transformed into numerical feature vectors using the TF-IDF representation. This technique captures the relative importance of words within each document while reducing the relevance of terms that are frequent across the entire corpus. TF-IDF works well with SVM because it produces a sparse, high-dimensional feature space, and SVM is particularly effective in this scenario. This enables the model to find an optimal decision boundary that separates documents according to their access level.

LSTM is a type of recurrent neural network that processes sequences of text tokens, capturing long-range dependencies through gating mechanisms designed to retain or discard information over time. The devised classification model receives word embeddings derived from the integer representations of tokens (based on the corpus vocabulary),

[2] https://www.nltk.org/

which are then fed into the LSTM layer. After processing the entire sequence of tokens, the final prediction is produced by a fully connected (FC) layer with a softmax activation function, which outputs a probability distribution over the document access levels. The access level with the highest probability is selected as the predicted class label.

In addition, the BiLSTM architecture extends the traditional LSTM by processing the input sequence in both forward and backward directions, allowing the model to capture contextual information from both past and future tokens simultaneously. The input processing remains the same as in the LSTM model. For prediction, the hidden states from the forward and backward directions for each token are concatenated and passed to the FC layer.

Finally, the fourth classification model is a pre-trained BERT, which captures deep contextual representations of language through attention mechanisms, trained with masked language modeling and next sentence prediction. For this model, the tokenization described in the previous section is not applied, as BERT uses its own subword-based tokenization strategy. Fine-tuning is performed on the entire model, and classification is achieved by adding a FC layer with a softmax activation function on top of the *[CLS]* token representation.

3.5.1. Classification Models Setup

To run the experiments, four classification pipelines were evaluated: TF-IDF combined with SVM, BERT, LSTM, and BiLSTM. Each model had its hyperparameters optimized to enhance predictive performance. This optimization step was conducted using a validation set.

Hyperparameter optimization for the SVM classification model was performed using grid-search cross-validation, which performs an exhaustive search over all possible hyperparameter combinations to identify the best configuration.

For the deep learning classification models, random search was employed to explore a limited number of randomly selected hyperparameter combinations. Furthermore, early stopping was applied to prevent overfitting by terminating training when the validation loss no longer improved. In our implementation, the monitored metric was the validation loss. Hyperparameter optimization was carried out using KerasTuner[3].

The evaluated hyperparameter values for each classification model are described below:

- **TF-IDF + SVM:**
  - Max features: 5000;
  - C: [0.1, 1, 10, 100, 1000];
  - Gamma: [1, 0.1, 0.01, 0.001, 0.0001];
  - Kernel: Radial Basis Function.
- **BERT:**
  - Base-Model: bert-base-multilingual-uncased;
  - Learning rate: [2.10^{-5}, 1.10^{-5}, 5.10^{-5}]

[3] https://keras.io/keras_tuner/

- **LSTM:**
  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Dropout rate: [0.1,0.2,0.3];
  - Learning rate: [1.10^{-2}, 1.10^{-3}, 5.10^{-3}]
- **BiLSTM**
  - Number of units: [128, 160, 192, 224, 256, 288, 320];
  - Dropout rate: [0.1,0.2,0.3];
  - Learning rate: [1.10^{-2}, 1.10^{-3}, 5.10^{-3}]

3.5.2. Stratified K-Fold Cross Validation

The evaluation method employed was Stratified K-Fold Cross-Validation with $K = 5$ folds. In each fold, the data was split into training (80%) and validation (20%) subsets, maintaining the original class distribution. This approach was selected to ensure that both training and validation sets are representative of the class proportions. During each iteration, the classification model is trained on the training data, and its hyperparameters are optimized using the validation set.

3.5.3. Evaluation Metrics

The F1-Score was chosen as the evaluation metric, as it represents the harmonic mean of precision and recall. Given the class imbalance observed in the corpus, the macro F1-Score was used to assess the overall performance of each classification model. The macro F1-Score is computed as the unweighted average of the single F1-Scores of the three class labels, assigning equal importance to each class regardless of the number of samples.

4. Experiments and Results

In this section, we present the experiments and analyze the results obtained to demonstrate the effectiveness of the proposed methodology. The source code, the SEI form templates used during corpus construction, and the final corpus are available in our public repository. [4].

Classification experiments were conducted to evaluate the corpus and determine the most suitable model for classifying the documents' privacy level. The experiments were performed on Google Colaboratory, and the pipeline was implemented in Python using Pandas[5], Numpy[6], NLTK[7], TensorFlow-Keras[8], PyTorch[9], and Scikit-Learn[10]. The models were trained using the Adam optimization algorithm [Kingma and Ba 2014].

Table 2 presents the results of the experiments in terms of F1-Score per class obtained for each of the selected models. The F1-Scores were calculated as the mean and

[4] https://gitlab.com/gvic-unb/classification-access-level-sei
[5] https://pandas.pydata.org/
[6] https://numpy.org/
[7] https://www.nltk.org/
[8] https://www.tensorflow.org/
[9] https://pytorch.org/
[10] https://scikit-learn.org/stable/

standard deviation across the five folds. The last column shows the Global Average (GA), calculated as the mean and standard deviation of the F1-Score macro results across the folds. The Support (Average) row represents the number of samples used for testing each access level and the overall GA.

It is worth noting that the lowest F1-Score values are associated with the Confidential class across all the models used. These documents are less frequent in the corpus compared to the others, which means the models had less data to learn from. Therefore, it is expected that the F1-Score for this document type would be lower. Given this, one possible way to improve these results is to increase the number of Confidential documents in the corpus. In contrast, the Public class yielded the highest F1-Score values across the four models used. This class label was the most frequent in the constructed corpus, meaning more documents were available for training the models.

Finally, based on the results, TF-IDF+SVM and BiLSTM showed superior performance across all document privacy levels, making them the most suitable models for the privacy level classification. Moreover, the F1-Score analysis suggests that the results are strongly influenced by the amount of training data available for each label.

Table 2. Classification results using Stratified K-Fold Cross Validation $(\mathbf{K = 5})$: mean and standard deviation of Macro F1-Score per class label across all folds.

Model	Public	Restrict	Confidential	GA
TF-IDF + SVM	0.99±0.01	0.98±0.02	0.91±0.05	0.96±0.02
BERT	0.99±0.01	0.96±0.01	0.81±0.09	0.92±0.04
LSTM	0.97±0.02	0.96±0.01	0.68±0.13	0.87±0.05
BiLSTM	1.00±0.00	0.98±0.01	0.90±0.11	0.96±0.04
Support (Average)	57.6	55.4	8.2	121.2

5. Conclusion

This paper presented an automatic approach for classifying the access level of documents from SEI according to public, restricted, and confidential. As this task is supervised, the lack of public corpora motivated us to construct a corpus constituted by public SEI documents collected from the Internet, while others were synthetically generated using ChatGPT. Finally, experiments were conducted to evaluate the performance of the classification models on the constructed corpus and including state-of-the-art models such as: SVM, LSTM, BiLSTM, and BERT.

The experiments demonstrated that SVM and BiLSTM achieved the highest macro F1-Scores, with performance varying according to the access level. The analysis of results by class label indicates that public documents yielded the best performance, followed by restricted documents. However, confidential documents showed the most significant discrepancies across models, probably due to lower representation in the corpus. Moreover, model performance was affected in the confidential access level, likely due to textual similarities with restricted documents.

For future work, we emphasize the importance of expanding the corpus by incorporating documents (1) of different types, (2) from various educational institutions and

other public organizations. Moreover, a multimodal approach could add significant value, as SEI supports multiple file formats such as images, PDFs, videos, and others.

Acknowledgments

This research was funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under Process No. 800019/2024-5, and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 88887.967792/2024-00. Ana Clara B. Borges acknowledges financial support from Grupo Mulheres do Brasil for her participation in the event.

References

Aldeen, Y. A. A. S., Salleh, M., and Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, 4(1):694.

Alparslan, E., Karahoca, A., and Bahşi, H. (2011). Classification of confidential documents by using adaptive neurofuzzy inference systems. *Procedia Computer Science*, 3:1412–1417. World Conference on Information Technology.

Bayer, M., Kaufhold, M.-A., and Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Cai, X., Xiao, M., Ning, Z., and Zhou, Y. (2023). Resolving the imbalance issue in hierarchical disciplinary topic inference via llm-based data augmentation. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1424–1429. IEEE.

Camarinha, A., Abreu, A., Júnior, M., and Cardoso, I. (2023). Users' perception of satisfaction of the eletronic information system – sei in the instituto federal de rondônia. *Journal of Information Systems Engineering and Management*, 8:18354.

Cloutier, N. A. and Japkowicz, N. (2023). Fine-tuned generative llm oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5181–5186. IEEE.

Coelho, G. M., Ramos, A. C., de Sousa, J., Cavaliere, M., de Lima, M. J., Mangeth, A., Frajhof, I. Z., Cury, C., and Casanova, M. A. (2022). Text classification in the brazilian legal domain. In *ICEIS (1)*, pages 355–363.

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., and Orosz, T. (2021). Challenges and open problems of legal document anonymization. *Symmetry*, 13(8):1490.

De Araujo, P. H. L., de Campos, T. E., Braz, F. A., and da Silva, N. C. (2020). Victor: a dataset for brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ghann, P., Tetteh, E. D., Asare Obeng, K., and Elias, M. (2022). Preserving the privacy of sensitive data using bit-coded-sensitive algorithm (bcsa). *International Journal of Recent Contributions from Engineering, Science amp; IT (iJES)*, 10(04):pp. 4–16.

Gottschalg-Duque, C. (2024). Towards the use of blockchain technology in sei, a brazilian electronic document and process management tool. In *Anais do II Colóquio em Blockchain e Web Descentralizada*, pages 26–31, Porto Alegre, RS, Brasil. SBC.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer Berlin Heidelberg.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Juez-Hernandez, R., Quijano-Sánchez, L., Liberatore, F., and Gómez, J. (2023). Agora: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. *Applied Soft Computing*, 145:110540.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., and Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Sangaroonsilp, P., Choetkiertikul, M., Dam, H. K., and Ghose, A. (2023a). An empirical study of automated privacy requirements classification in issue reports. *Automated Software Engineering*, 30(2):20.

Sangaroonsilp, P., Dam, H. K., Choetkiertikul, M., Ragkhitwetsagul, C., and Ghose, A. (2023b). A taxonomy for mining and classifying privacy requirements in issue reports. *Information and Software Technology*, 157:107162.

Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Sousa, S. and Kern, R. (2023). How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, 56(2):1427–1492.

Sulavko, A., Varkentin, Y., Panfilova, I., and Samotuga, A. (2024). Automatic classification of text messages by confidentiality level based on ensemble of artificial neural networks. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 484–492. IEEE.

Wu, J. M.-T., Srivastava, G., Jolfaei, A., Fournier-Viger, P., and Lin, J. C.-W. (2021). Hiding sensitive information in ehealth datasets. *Future Generation Computer Systems*, 117:169–180.