Classification of Irregularity Communications in Public Ombudsmen Using Supervised Learning Algorithms

Fábio Cordeiro[1,2], **Ricardo de Andrade Lira Rabelo**[1], **Raimundo Santos Moura**[1]

[1]Universidade Federal do Piauí – UFPI
Departamento de Computação
Teresina – PI – Brasil

[2]Tribunal de Contas do Estado do Piauí
Teresina – PI – Brasil

fabiocordeiro@gmail.com, {ricardoalr, rsm}@ufpi.edu.br

Abstract. *The objective of this work is to evaluate Supervised Learning algorithms in the task of classifying irregularities in Public Ombudsman Offices of Courts of Auditors. We intend to contribute effectively to improving the analysis of these communications, enabling a faster response to the citizen. Due to the imbalance of the original releases, we apply data resizing techniques before training the models. Classical ML algorithms (Naive Bayes, Decision Tree, Random Forest, K Nearest Neighbor, and Support Vector Machine) were compared with the Deep Learning Bidirectional Encoder Representations from Transformers (BERT) model and variations of text representation with Word Embeddings. The best results were obtained by the BERT model with the resampling dataset, reaching 96% in the F1-Score metric.*

Resumo. *O objetivo deste trabalho é avaliar modelos de Aprendizado de Máquina (AM) na tarefa de classificação de comunicados de irregularidades em Ouvidorias Públicas de Tribunais de Contas. De maneira geral, pretende-se contribuir de forma efetiva para melhorar a triagem desses comunicados, possibilitando maior celeridade na resposta ao cidadão. Devido ao desbalanceamento do* dataset *original, foram aplicadas técnicas de redimensionamento de dados antes da etapa de treinamento dos modelos. Algoritmos clássicos de* Machine Learning *(Naive Bayes, Decision Tree, Random Forest, K Nearest Neighbor e Support Vector Machine) foram comparados com o modelo de* Deep Learning *Bidirectional Encoder Representations from Transformers (BERT) e variações de representação dos textos com* Word Embeddings. *Os melhores resultados foram obtidos pelo modelo BERT com o dataset redimensionado, atingindo 96% na métrica F1-Score.*

1. Introdução

Previsto na Constituição Federal de 1988, o controle social é um ato do exercício da cidadania[1], contribuindo para o fortalecimento da soberania popular. Dessa forma, a efetiva participação do cidadão, por meio de ferramentas como dispositivos legais e tecnológicos, é cada vez mais estimulada pelos órgãos de controle.

[1]Segundo o parágrafo único do art. 1º da Constituição Federal de 1988, todo o poder emana do povo, que o exerce por meio de representantes eleitos ou diretamente. (Grifo nosso).

A Lei n° 12.527/2011, conhecida como Lei de Acesso à Informação e a Lei n° 13.460/2017, que dispõe sobre a participação, proteção e defesa dos direitos dos usuários de serviços públicos, são exemplos de mecanismos criados para incentivar e garantir ao cidadão o exercício do controle social. Segundo [Ouvidoria Geral da União 2018], as possíveis manifestações descritas nessas leis são: pedidos de informação (regida pela Lei de Acesso à Informação), denúncia (ou comunicação de irregularidade), reclamação, sugestão, elogio e solicitação.

No setor público, de acordo com [Ouvidoria Geral da União 2018], o canal de comunicação entre o cidadão e a Administração Pública é feito, geralmente, por meio das Ouvidorias. Esses órgãos são responsáveis pelo recebimento e tratamento das reclamações, solicitações, comunicações de irregularidades (denúncias de ouvidoria) e sugestões feitas pelos cidadãos referentes às ações da gestão pública, aplicação de recursos, prestação de serviços, ou qualquer função exercida pela Administração Pública.

No âmbito dos Tribunais de Contas, além das demandas internas decorrentes de obrigações constitucionais, como a análise das contas dos gestores públicos, observa-se o crescimento de demandas externas e imprevisíveis advindas da sociedade, com potencial de comprometer o cronograma das atividades planejadas pela governança da instituição, dificultando a resposta aos cidadãos de forma célere.

Além de receber comunicados de irregularidades, as Ouvidorias Públicas servem para conhecer as demandas da população, por meio do recebimento de manifestações que podem ser reclamações, solicitações, sugestões e elogios, relativos aos serviços e políticas públicas. A Ouvidoria do Tribunal de Contas do Estado do Piauí (TCE/PI) recebeu em 2015, por meio de formulário Web, 589 manifestações. Em 2021, foram recebidos 2.609 comunicados, mais de 4 vezes do que foi recebido em 2015. O aumento dessa demanda exige uma contrapartida na estruturação das Ouvidorias e setores internos, que são acionados nos casos em que a própria Ouvidoria não tenha subsídios para dar uma resposta ao cidadão.

A análise inicial dessas manifestações consiste no processo de triagem, que verifica a pertinência e a materialidade da comunicação. Se o responsável pela triagem possuir elementos suficientes, ele pode responder a comunicação instantaneamente. Porém, se for necessária uma análise mais aprofundada, ele pode encaminhar para outros setores internos competentes, como as diretorias técnicas do TCE, corregedorias e controladorias internas de seus jurisdicionados ou, ainda, para outras instituições de controle, como o Ministério Público.

Ao analisar cada uma das demandas recebidas, é importante fazer uma leitura minuciosa para descobrir do que se trata e, posteriormente, fazer a classificação do assunto a que se refere a manifestação, pois a destinação correta depende do teor da solicitação.

Desde 2015, a Ouvidoria do TCE/PI disponibiliza um formulário eletrônico para receber manifestações de diversos assuntos. Com relação ao assunto Comunicados de Irregularidades, objeto de pesquisa deste trabalho, foram recebidas quantidades crescentes de manifestações, como pode ser observado na Figura 1.

É importante notar o aumento crescente do interesse popular em participar no controle das atividades da Administração Pública. Porém, em 2020, por conta da pandemia, houve uma redução no número de comunicações recebidas.

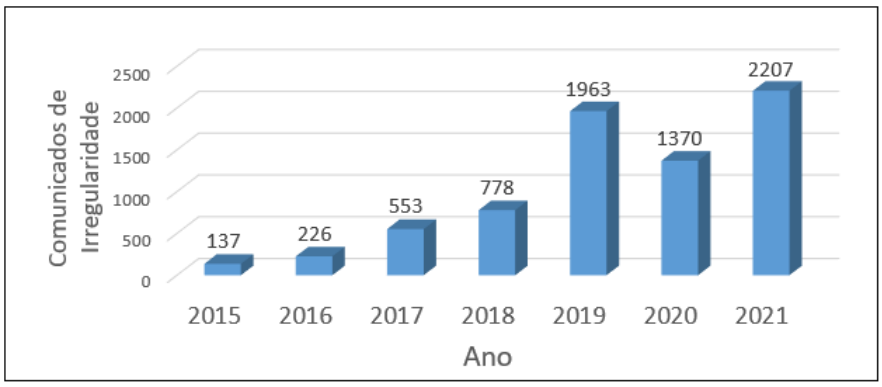

Figura 1. Comunicados de Irregularidade recebidos por meio da Ouvidoria.

O formulário atual da Ouvidoria do TCE/PI permite que o cidadão efetue 5 (cinco) tipos diferentes de manifestação: Comunicação de Irregularidades, Sugestão, Reclamação, Pedido de Informação e Elogio. As manifestações podem ser referentes ao próprio TCE/PI ou aos órgãos jurisdicionados, que são fiscalizados pelo TCE. Os Comunicados de Irregularidades, em geral, são relacionados aos entes fiscalizados.

Especificamente, sobre os comunicados de irregularidades, o formulário disponibiliza uma lista contendo 9 (nove) assuntos: Acesso à informação, Acumulação de cargos, Atraso salarial, Concurso e admissão, Licitação, Nepotismo, Previdência, Transparência e Outros. Além da quantidade de assuntos do formulário não atender a atual necessidade, foram detectados muitos erros provenientes da classificação realizada pelo cidadão, o que dificultou o aproveitamento dessas informações.

Devido ao aumento da participação popular na gestão pública e à deficiência do atual sistema em fazer uma correta triagem desses comunicados de irregularidades foi desenvolvido um novo sistema para a Ouvidoria do TCE/PI com uma nova classificação de assuntos, considerando mais de 90 (noventa) classes, organizadas hierarquicamente. Atualmente, este sistema está em fase de implantação. A nova classificação demandará uma análise mais detalhada de cada comunicado e, consequentemente, o aumento do custo hora-homem, já que a nova classificação será feita por funcionários da Ouvidoria do TCE/PI.

O objetivo deste trabalho é avaliar modelos de Aprendizado de Máquina Supervisionado (AM) na tarefa de classificação de comunicados de irregularidades, recebidos por meio da Ouvidoria do TCE/PI. Como resultado esperado pretende-se contribuir de forma efetiva na melhoria da triagem desses comunicados, possibilitando maior celeridade na resposta ao cidadão.

O restante deste artigo está organizado da seguinte maneira: a Seção 2 apresenta os principais trabalhos relacionados com a tarefa de classificação de informações textuais. A Seção 3 descreve uma visão geral da metodologia utilizada para o desenvolvimento deste trabalho e comenta as quatro primeiras etapas. A seguir, a Seção 4 discute detalhes da etapa de criação dos modelos com os dados originais e com o uso de técnicas de redimensionamento. A Seção 5 apresenta e discute os resultados dos modelos avaliados, considerando três grupos: i) modelos clássicos; ii) SVMs com diferentes formas de representação de *embeddings*; e iii) modelo de língua BERT [Devlin et al. 2018]. Por

fim, a Seção 6 conclui o artigo e aponta alguns trabalhos futuros.

2. Trabalhos Relacionados

As buscas por publicações relacionadas ao presente trabalho foram feitas por meio de pesquisas por palavras chaves no Google Acadêmico, na biblioteca SOL da SBC e em periódicos CAPES. Existem diversos trabalhos relacionados à área de PLN, desde classificadores de texto e tradutores automáticos, até assistentes virtuais e chatbots. No entanto, poucos trabalhos são diretamente relacionados ao objeto deste estudo. A Dissertação de Mestrado de Patrícia Andrade [Andrade 2015] possui como objetivo principal classificar "Denúncias e Manifestações" do portal de Ouvidoria da Controladoria Geral da União CGU. Ela trata o problema de classificação multiclasse com 64 classes distintas. Ela usou representação textual TF-IDF e avaliou os algoritmos: *Support Vector Machine* (SVM), *Random Forest*, *Naïve Bayes*, Árvore de decisão e a Codificação adaptativa de Huffman em conjunto com o Minimum Description Length (CAH+MDL). Os melhores resultados foram obtidos utilizando o algoritmo CAH+MDL e a classificação multi-label, com precisão de 84%. Destaca-se que este trabalho foi utilizado como base para a estratégia de busca *snowballing*, que resultou na escolha de mais 4 (quatro) trabalhos. Adicionalmente, outros dois trabalhos foram analisados por também estarem relacionados ao presente estudo.

O trabalho de [Rocha 2019] realiza a classificação de documentos jurídicos, no qual as classes são relacionadas à pedidos de ações trabalhistas. Trata-se de classificação multiclasse com 36 assuntos organizados de forma hierárquica com 5 níveis. O modelo de representação textual utilizado foi o TF-IDF, considerando termos que aparecem em pelo menos 5 documentos e ignorando termos que aparecem em mais de 80% dos documentos. Foram analisados os algoritmos: SVM, *Random Forest*, *Multinomial Naïve Bayes* e a Rede Neural *Multilayer Perceptron* (MLP). O melhor resultado foi obtido com o modelo MLP com precisão de 46,04%.

[Palma et al. 2021], tratam as solicitações de informações do cidadão endereçadas ao Ministério das Minas e Energia do Brasil como um problema de classificação multiclasse com 44 classes, usando aprendizagem de máquina não supervisionada. Os autores usaram representação textual TF-IDF, considerando termos presentes em pelo menos 15 documentos e ignorando termos que aparecem em mais de 50% das solicitações. Foram analisados diversos algoritmos de AM, sendo o melhor desempenho obtido com o algoritmo XGBoost, 75% de acurácia.

[Souza 2021] apresenta um estudo comparativo entre algoritmos clássicos de AM e o modelo de *Deep learning* BERT (*Bidirectional Encoder Representations from Transformers*). Ele trata a tarefa de classificação multiclasse com 424 classes de processos jurídicos no âmbito da Procuradoria Geral do Distrito Federal (PGDF). Como as classes são desbalanceadas, foram analisadas algumas técnicas de balanceamentos, a saber: *SMOTE*, *Random Under-Sampling* e *SMOTEENN*. Após vários testes dos algoritmos, incluindo o algoritmo LDA para classificação não supervisionada, a quantidade de classes foi reduzida. Com a redução das classes e com o processo de balanceamento artificial dos dados, o modelo BERT alcançou 89% de precisão e o algoritmo *Gradient Boosting* obteve 90% de precisão.

[Gusmão et al. 2021] trabalharam com a classificação multiclasse de denúncias

criminais efetuadas por meio do aplicativo Disque Denúncia do Rio de Janeiro. Eles consideraram 15 classes, representação textual TF-IDF e avaliaram apenas o algoritmo SVM. Foram analisadas diferentes técnicas de pre-processamento, como: conversão dos caracteres para minúsculas, remoção de números, sinais de pontuação, substituição de acrônimos, remoção de *stopwords*. Além disse, eles avaliaram o impacto de usar correção ortográfica e *stemming* das palavras. O melhor resultado foi obtido com a aplicação de todas as técnicas de pré-processamento, correção ortográfica e o processo de *stemming*, atingindo 76,11% de acurácia.

[Lee et al. 2022] avaliaram 7 (sete) variações pré-treinadas do BERT para tarefas de compreensão de linguagem natural. Eles investigaram a maneira como os modelos classificam a tecnologia climática em propostas de pesquisa escritas no idioma coreano, abrangendo 45 classes diferentes. O resultado mostrou que o modelo *KLUE-BERT*, um modelo BERT pré-treinado com um grande e diversificado *corpus* coreano) superou os demais modelos, alcançando 72% na métrica Macro-F1.

[Tang et al. 2021] apresentam uma pesquisa sobre classificação de reclamações de clientes de concessionárias de energia elétrica, em *corpus* desbalanceados. Trata-se de um problema de classificação muticlasse, com hierarquia de classes, no qual as classes mais altas (classes rasas) possuem muitas amostras e as classes mais baixas (classes profundas) possuem poucas amostras, caracterizando o desbalanceamento dos dados. O modelo BERT obteve 92,79% de acurácia para as classes rasas e, quando utilizado a representação vetorial com *embeddings word2vec*, o modelo obteve 73,5% de acurácia para as classes profundas.

A Tabela 1 sumariza algumas informações e os principais resultados dos trabalhos discutidos anteriormente. Por fim, é importante ressaltar que alguns trabalhos não apresentam todos resultados, o que impossibilitou a unificação em uma única métrica para fins de comparação.

Tabela 1. Resumo dos trabalhos relacionados

Trabalho	Representação vetorial	Data Augmentation	Algoritmo de Deep Learning	Melhor Resultado
[Andrade 2015]	TF-IDF	Não	Não	84% de precisão
[Rocha 2019]	TF-IDF	Não	MLP	46,01% de precisão
[Palma et al. 2021]	TF-IDF	Não	Não	75% de acurácia
[Souza 2021]	Bag-of-Words	SMOTEENN	BERT	95% de F1-score
[Gusmão et al. 2021]	TF-IDF	Não	Não	76,11% de acurácia
[Lee et al. 2022]	Embeddings	Não	BERT	72% de Macro-F1
[Tang et al. 2021]	word2vec	Back Translation e Substituição de Sinônimos	BERT	92,79% de acurácia

3. Metodologia

Para a construção dos modelos de AM Supervisionados utilizou-se uma metodologia baseada no *Pipeline* descrito por [Vajjala et al. 2020]. Além das etapas tradicionais, foram incluídas duas etapas extras: *Rotulagem* e *Redimensionamento*, como apresentada na Figura 2. Todas as etapas serão descritas a seguir:

Durante a etapa de *Coleta* foram obtidos 7.653 registros da base de dados do sistema atual da Ouvidoria. A etapa de *Rotulagem* foi realizada por dois especialistas, funcionários do TCE/PI com mais de 5 anos de experiência. A qualidade da rotulagem

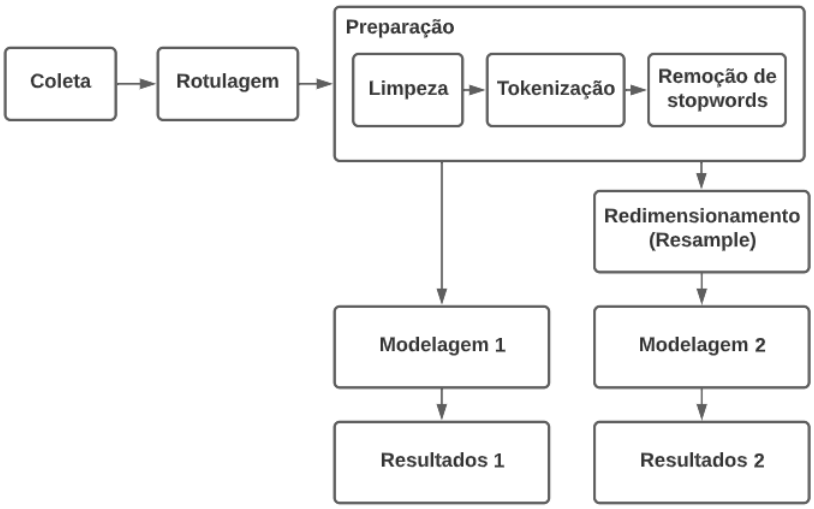

Figura 2. Adaptação do *Pipeline* de [Vajjala et al. 2020].

não foi aferida devido a necessidade de se rotular a maior quantidade possível de comunicados. Foram rotulados manualmente 463 registros, distribuídos entre 13 classes diferentes. Destaca-se que a classe Licitações e Contratos Públicos possui 166 registros, sendo a classe majoritária. Já a classe com menor quantidade foi Previdência Social com apenas 1 registro.

Devido à pouca quantidade de dados, foi utilizada uma linha de corte para as classes que apresentaram quantidade menor que 5 registros. São elas: Finanças públicas e orçamento, Previdência social e Segurança pública. Essas três classes foram reclassificadas para pertencerem à classe Outros.

A Tabela 2 mostra a classificação antes do corte. Após o corte e o processo de reclassificação, a classe Outros passou a ter 29 registros.

Tabela 2. Quantidade de comunicados por classe

Classe	Qtd
Licitações e contratos públicos	166
Transparência e acesso à informação	90
Remuneração, direitos e deveres dos agentes públicos	43
Acumulação de cargos	38
Concursos e admissões	37
Nepotismo	31
Outros	22
Serviços de educação	14
Controle interno	10
Serviços de saúde	5
Finanças públicas e orçamento	4
Segurança pública	2
Previdência social	1
Total Geral	**463**

A etapa de *Preparação* contou com a limpeza dos dados, tokenização dos textos e remoção de *stopwords*. A etapa de *Modelagem e avaliação* foi realizada de duas for-

mas diferentes: na primeira utilizou-se o resultado da etapa de *Preparação* sem nenhuma alteração; e a segunda forma realizou-se o processo de redimensionamento (*Resample*) dos dados antes da etapa de *Modelagem e avaliação*.

No redimensionamento foi utilizada a técnica *Undersampling*, que é a eliminação de dados de classes majoritárias. Essas classes, por conterem mais dados, podem enviesar os modelos. Para fazer o *Undersampling* foi estabelecido um teto máximo e o valor foi calculado a partir da quantidade de comunicados dividida pela quantidade de classes, o que resultou em 46,3, esse valor foi aproximado para 50, com isso, classes maiores foram reduzidas para 50 registros. Utilizou-se também a técnica *Oversampling*, que é o aumento de dados, criando-se amostras artificiais para as classes minoritárias. Ao final da etapa todas as classes totalizaram 50 registros cada. Para realizar o *Oversampling* foi utilizada a biblioteca python *NLPAUG*[2] específica tarefa de aumento de texto, juntamente com o modelo BERT pré-treinado para a língua portuguesa, *BERTimbau*[3].

É importante mencionar que a técnica de *Oversampling* sobre classes com poucos registros pode provocar inviesamento dos modelos, pois dados artificiais são gerados para o balanceamento das classes. Por exemplo, no nosso caso, as classes Serviço de Saúde, Controle Interno e Serviços de Educação originalmente possuíam 5, 10 e 14 registros, respectivamente, e com o *Oversampling* ficaram com 50 registros.

No entanto, para evitar esse enviesamento está sendo elaborado uma versão estendida do *corpus* de comunicados de irregularidade.

A Figura 3 apresenta o *dataset* original, antes do redimensionamento. Duas classes sofreram *Undersampling*: Licitações e contratos públicos e Transparência e acesso à informação. As demais classes sofreram aumento de dados até alcançar o total de 50 registros.

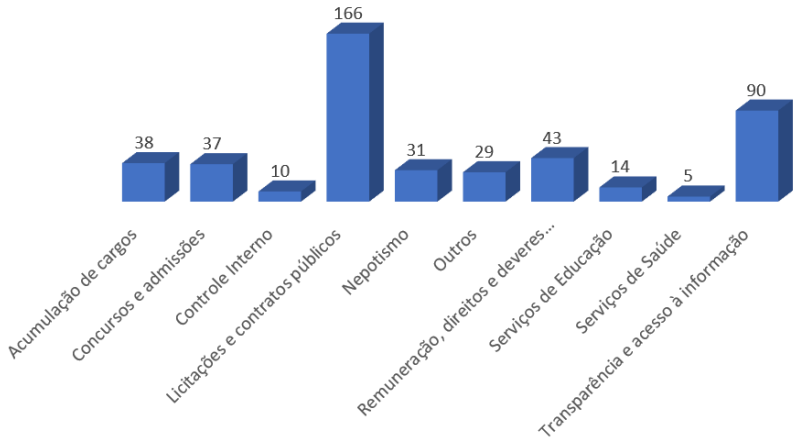

Figura 3. *Dataset* original, antes do redimensionamento.

4. Modelagem

O *dataset* original (*Modelagem 1*) e o *dataset* redimensionado (*Modelagem 2*), foram salvos em locais diferentes e, posteriormente separados em conjunto de treinamento, teste

[2] https://nlpaug.readthedocs.io/en/latest/
[3] https://huggingface.co/neuralmind/bert-base-portuguese-cased

e validação, utilizando a função *python train_test_val* [4].

O *dataset* original, por conter classes com apenas cinco registros, foi dividido na proporção 70% para treinamento, 15% para teste e 15% para validação, enquanto o *dataset* redimensionado foi dividido na proporção de 80% para treinamento, 10% para teste e 10% para validação.

A ferramenta usada para executar os modelos foi o Google Colab[5]. A modelagem dos algoritmos de AM foi dividida em três conjuntos: Algoritmos clássicos, *Support Vector Machine (SVM)* com *Word Embeddings* e Aprendizado profundo (BERT). Para execução dos modelos clássicos utilizou-se a biblioteca *scikit-learn*[6] da linguagem Python. Cada modelo foi avaliado tanto com o *dataset* original, quanto com o *dataset* redimensionado utilizando os parâmetros de configuração *default*. Nos modelos clássicos, os textos foram representados utilizando *Bag of Words* com pesos TF-IDF.

Na representação **word2vec** foram utilizadas as *Word Embeddings* do **NILC**[7], *CBOW* e *Skip-gram*, de 300 dimensões. Os modelos **doc2vec** foram gerados a partir dos dados de treinamento e teste de cada *dataset*.

Por fim, o modelo BERT foi implementado usando a biblioteca *Pytorch*[8]. A tarefa de ajuste fino foi feita utilizando o modelo BERTimbau, modelo BERT pré-treinado para o português. O resultado da implementação foi um classificador que possui uma camada BERT pré-treinada (BERTimbau), uma camada de *dropout* para evitar o *overfitting*, e para a camada de saída do classificador é uma camada linear. Na camada linear de saída foi passado o valor *default* para parâmetro *in_features* que representa o tamanho do vetor da camada oculta (*hidden_size*), que é de 768. O parâmetro *out_features* da camada linear foi setado com o valor 10, que corresponde ao número de classes do treinamento do ajuste fine. Os valores dos hiperparâmetros do modelo foram os seguintes:

- modelo pré-treinado para o português *BERTimbau*, *base* e *large*;
- comprimento máximo do tensor (max_length = 128);
- quantidade de dados por *batch* (batch_size = 32);
- quantidade de subprocessos do *dataloader* (num_workers = 2);
- quantidade de épocas para treinamento do modelo (epochs = 10);
- taxa de *dropout* (dropout = 0,10);
- função de otimização *AdamW*;
- taxa de aprendizagem (lr = $5e^{-5}$) - *learning rate*;
- fator epsilon (eps = $1e^{-8}$) - termo adicionado ao denominador da função de otimização para evitar a divisão por zero;
- quantidade de classes igual a 10.

Durante os testes, os hiperparâmetros foram variados em diversas combinações e a cada nova escolha dos parâmetros o modelo foi retreinado, anotando-se os resultados. O melhor resultado obtido foi escolhido e os hiperparâmetros fixados. Observou-se que valor *default* 0.10 foi o melhor ajuste para o parâmetro *dropout*. A função de otimização

[4] https://stackoverflow.com/questions/70911657/data-subset-in-python
[5] https://colab.research.google.com/
[6] https://scikit-learn.org/
[7] http://www.nilc.icmc.usp.br/embeddings
[8] urlhttps://pytorch.org/

AdamW foi escolhida por ser a recomendada para trabalhos de classificação multiclasse. O parâmetro Taxa de Aprendizagem, que serve para encontrar o mínimo global do gradiente, foi aleatoriamente testado com vários valores, até chegar ao valor de $5e^{-5}$, que melhor contribuiu para o bom resultado. O parâmetro *eps* permaneceu com seu valor *default*.

Devido à baixa quantidade de dados anotados e o desbalanceamento das classes os primeiros resultados da modelagem foram ruins. O alto custo para fazer anotação do corpus levou a adoção da geração artificial de dados anotados, por meio de técnicas de redimensionamento. Durante a modelagem, ficou evidente a importância do uso das técnicas de redimensionamento das amostras para melhorar o resultado dos modelos, como será discutido na próxima seção.

5. Resultados

Os resultados dos modelos analisados com testes sem redimensionamento e com redimensionamentos são apresentados na Tabela 3, que contém as médias das medidas de acurácia, precisão, revocação e f1-score. Para facilitar o entendimento, os modelos foram divididos em 3 grupos: i) modelos clássicos; ii) modelo SVM com variadas formas de representação das *Embeddings*; e iii) modelo BERT.

Tabela 3. Resultados dos Testes

Grupo	Modelo	Rep. Vetorial	Métricas (média)			
			Acurácia	Precisão	Revocação	F1-score
Resultados 1 - sem redimensionamento						
(i)	Naive Bayes	TF-IDF	0,47	0,40	0,47	0,36
	Logistic Regression	TF-IDF	0,69	0,74	0,69	0,63
	SVM[a]	**TF-IDF**	**0,84**	**0,83**	**0,84**	**0,83**
	Decision Tree	TF-IDF	0,69	0,68	0,69	0,67
	Random Forest	TF-IDF	0,76	0,83	0,76	0,75
	K Nearest Neighbor	TF-IDF	0,80	0,80	0,80	0,79
(ii)	**SVM**	**TF**[b]	**0,83**	**0,74**	**0,68**	**0,69**
	SVM	CBOW	0,51	0,27	0,24	0,23
	SVM	Skip-gram	0,43	0,19	0,17	0,16
	SVM	DM	0,67	0,44	0,42	0,42
	SVM	DBOW	0,67	0,45	0,52	0,46
(iii)	**$BERT_{BASE}$**	**Contextual Embeddings**[c]	**0,86**	**0,70**	**0,67**	**0,67**
Resultados 2 - com redimensionamento						
(i)	Naive Bayes	TF-IDF	0,84	0,88	0,84	0,83
	Logistic Regression	TF-IDF	0,86	0,86	0,86	0,85
	SVM	TF-IDF	0,86	0,87	0,86	0,85
	Decision Tree	TF-IDF	0,74	0,77	0,74	0,73
	Random Forest	**TF-IDF**	**0,88**	**0,89**	**0,88**	**0,88**
	K Nearest Neighbor	TF-IDF	0,74	0,76	0,74	0,72
(ii)	SVM	TF	0,68	0,71	0,68	0,65
	SVM	CBOW	0,32	0,39	0,32	0,28
	SVM	Skip-gram	0,22	0,12	0,22	0,14
	SVM	DM	0,64	0,68	0,64	0,64
	SVM	**DBOW**	**0,80**	**0,84**	**0,80**	**0,78**
(iii)	**$BERT_{BASE}$**	**Contextual Embeddings**	**0,96**	**0,97**	**0,96**	**0,96**

[a] Support Vector Machine; [b] Termo Frequência (bag-of-words) ;[c] Embeddings contextualizadas

Nos testes com o *dataset* sem redimensionamento o melhor resultado dos modelos clássicos foi o modelo SVM, com acurácia média de 0,84. O pior modelo foi o Naive Bayes, com 0,47 de acurácia média. Entre as representações de *Word Embedding* o melhor resultado ficou com a representação Termo Frequencia (TF), com acurácia média de 0,83,

porém, a medida de F1-Score, considerada melhor para avaliar conjuntos desbalanceados, resultou em 0,69. As representações word2vec (CBoW e Skip-Gram), mesmo usando as *Embeddings* pré-treinadas do NILC, foram as que mostraram os piores resultados, com destaque para o 0,43 de acurácia e 0,16 de F1-Score da representação word2vec - Skip-Gram. O modelo BERT superou o modelo SVM na medida acurácia, porém foi inferior nas outras medidas, sugerindo que o viés das classes majoritárias influenciaram mais o modelo BERT que o modelo SVM.

Os resultados com redimensionamento apresentaram menor variação entre as medidas. O melhor resultado entre os modelos clássicos foi o modelo *Random Forest*, com 0,88 de acurácia média. Nas representações de *Embeddings*, o primeiro lugar ficou com o modelo doc2Vec - DBoW, com 0,80 de acurácia. Destaca-se que as representações word2vec (CBoW e Skip-Gram) permaneceram nas piores colocações. Após várias tentativas de ajustar os hiperparâmetros do modelo BERT, obteve-se o melhor resultado de todos os modelos nas duas situações, confirmando a capacidade do modelo BERT em tarefas de PLN. O BERT atingiu 0,96 em quase todas as medidas com o *dataset* redimensionado.

As figuras 4 e 5 mostram as matrizes de confusão para o modelo SVM e para o modelo BERT, respectivamente, considerando os dados sem e com redimensionamento (*resampling*).

Na matriz do lado direito da Figura 4, é possível perceber que a maioria das classes tiveram acerto de 100%, enquanto que do lado esquerdo, somente a classe majoritária Licitações e Contratos Públicos teve 100% de acerto.

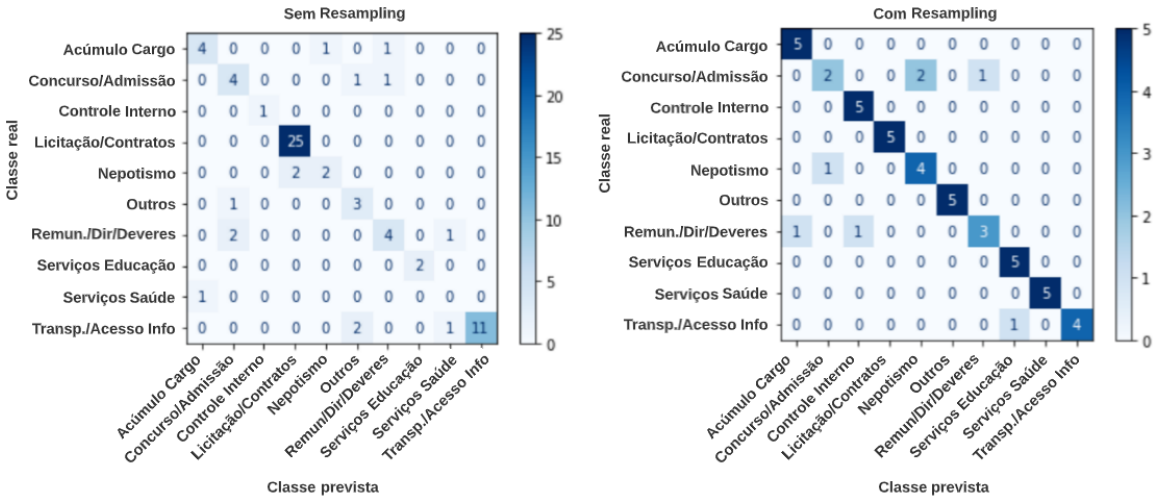

Figura 4. Modelo SVM - Antes e depois do redimensionamento

Com relação ao modelo BERT (ver Figura 5), obeserva-se a grande diferença entre os modelos: com o *dataset* sem redimensionamento, o BERT acerta 100% da classe Licitações e contratos públicos, mas erra em quase todas as outas classes. Já com o *dataset* com redimensionamento (lado direito da Figura 5), o BERT só errou na classe Transparência e acesso à informação, o que faz sentido, pois essa classe pode estar relacionada a falta de transparência em diversos assuntos, como saúde, educação, segurança, etc.

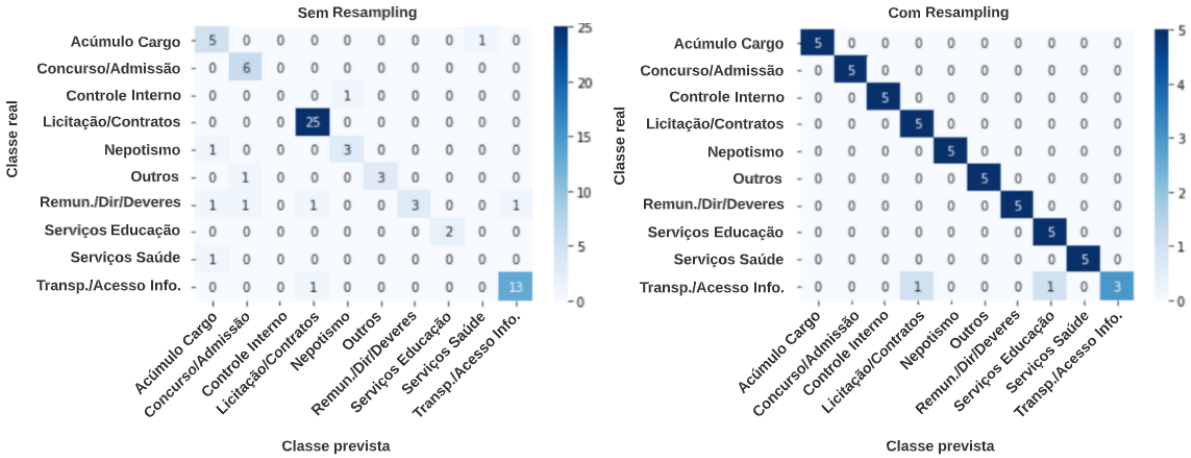

Figura 5. Modelo *BERT* - Antes e depois do redimensionamento

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou um estudo de modelos clássicos de aprendizagem supervisionada, modelos SVM com uso de representação de *Embeddings* e modelos de língua BERT para a tarefa de classificação de comunicados de irregularidades do TCE/PI. Os modelos foram investigados, considerando o *dataset* original e com redimensionamento dos dados, por meio de técnicas de *undersampling* e *oversampling*. Os resultados obtidos após o redimensionamento dos dados são considerados promissores, mas novos experimentos são necessários, devido à pequena quantidade de dados avaliados.

A análise os resultados mostra que alguns modelos apresentaram melhorias significativas após o redimensionamento dos dados, e outros modelos não. Observando a métrica F1-score, somente três modelos não obtiveram ganhos após o redimensionamento: *K-Nearest Neighbor*, *SVM-TF* e *SVM-Skip-gram*. O modelo que obteve maior aumento em suas medidas de performance foi o *Naïve Bayes*, seguido do modelo BERT. A Tabela 4 mostra as medidas de performance do modelo BERT, antes e depois do redimensionamento.

Tabela 4. Medidas do Modelo BERT

Redimensionamento	Acurácia	Precisão	Revocação	F1-score
SEM	0,86	0,70	0,67	0,67
COM	0,96	0,97	0,96	0,96

Adicionalmente, os resultados obtidos indicam a importância da etapa de redimensionamento de dados no processo de classificação de comunicados de irregularidades, pois o ganho de performance dos modelos analisados justifica a adoção desta etapa. Porém, a versão estendida do *corpus* deve ser utilizada para evitar o enviesamento provocado pelas classes minoritárias. Outro ganho diz respeito a redução dos custos de anotação manual dos dados.

Como contribuição prática, o desenvolvido pleno deste projeto permitirá a diminuição do custo de classificação dos comunicados, dando maior celeridade e eficiência aos trabalhos da Ouvidoria.

Como trabalhos futuros destaca-se: i) criar a versão estendida do *corpus* de comunicados anotados com o maior número de registros possível; ii) avaliar outros modelos de *Deep Learning*, por exemplo o *fast.ai*, para uma comparação com o modelo BERT; iii) fazer a interpretabilidade dos resultados; e iv) desenvolver um *web service* para receber o texto dos comunicados de irregularidades e devolver a lista de possíveis classes.

7. Agradecimentos

Os autores agradecem ao TCE/PI pelo total apoio e esperamos que este trabalho contribua para o melhor desempenho de sua Ouvidoria.

Referências

Andrade, P. H. M. A. d. (2015). Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na cgu. Dissertação de mestrado, Universidade de Brasília.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gusmão, C., Figueiredo, K., and Brito, W. A. (2021). Técnicas de processamento de linguagem natural em denúncias criminais: Automatização e classificação de texto em português coloquial. In *Anais do XLVIII Seminário Integrado de Software e Hardware*, pages 172–182. SBC.

Lee, E., Lee, C., and Ahn, S. (2022). Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9):4522.

Ouvidoria Geral da União (2018). *Manual de Ouvidoria Pública. https://repositorio.cgu.gov.br/bitstream/1/29959/14/manual_de_ouvidoria_publica.pdf*.

Palma, I., Ladeira, M., and Reis, A. C. (2021). Machine learning predictive model for the passive transparency at the brazilian ministry of mines and energy. pages 76–81. Association for Computing Machinery.

Rocha, A. C. P. (2019). Mineração de textos para classificação de processo judiciais trabalhistas. Dissertação de mestrado, Universidade de Brasília.

Souza, R. C. (2021). Uma comparação entre métodos e classificadores em documentos jurídicos de atividades processuais repetitivas na PGDF. Dissertação de mestrado, Universidade de Brasília.

Tang, X., Mou, H., Liu, J., and Du, X. (2021). Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports*, 11(1):1–11.

Vajjala, S., Majumder, B., Gupta, A., and Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. O'Reilly.