Sumarização de Denúncias: Proposta e Avaliação de Métodos de Geração de Resumos

Eduardo de Paiva[1,2], **Fernando Sola Pereira**[1], **Nelson Ebecken**[2]

[1]Diretoria de Informações Estratégicas – Controladoria Geral da União (CGU)
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro, Brasília – DF - Brasil

[2]COPPE – Universidade Federal do Rio de Janeiro (UFRJ)
Av. Horácio Macedo, 230, Centro de Tecnologia, 21941-909, Rio de Janeiro –RJ – Brasil

{eduardo.paiva,fernando.pereira}@cgu.gov.br, nelson@ntt.ufrj.br

Abstract. *The Brazilian legal system allows any citizen to report irregularities that are happening in the Public Administration. However, the volume of information present in the complaints' texts makes their treatment very expensive. Thus, there is a need for summarization methods capable of summarizing the complaints' texts. This paper's goal is to propose and evaluate two strategies for summarizing complaints: one based on the BERT language model and the other on word frequency. The study concluded that, for the purpose in question, the summaries generated by the BERT model were better than those generated by word frequency.*

Resumo. *O ordenamento jurídico brasileiro permite que qualquer cidadão faça denúncias sobre irregularidades que estejam acontecendo na Administração Pública. No entanto, o volume de informações presentes nos textos das denúncias torna o seu tratamento muito custoso. Dessa forma, surge a necessidade de métodos de sumarização capazes de resumir os textos das denúncias. O objetivo desse artigo é propor e avaliar duas estratégias de sumarização de denúncias: uma baseada no modelo de linguagem BERT e outra em frequência de palavras. O estudo concluiu que, para o propósito em questão, os resumos gerados pelo modelo BERT eram melhores que os gerados pela frequência de palavras.*

1. Introdução

O ordenamento jurídico brasileiro permite que qualquer cidadão faça denúncias sobre irregularidades que estejam acontecendo na Administração Pública.

No entanto, para que as denúncias possam ser apuradas, elas precisam reunir informações consistentes sobre os fatos narrados. Logo, a primeira atividade do tratamento das denúncias é chamada de análise de aptidão e tem o objetivo de classificar, com base em todo o material recebido, se uma denúncia deve ser considerada como apta ou não.

Dessa forma, a análise de aptidão de denúncias é um problema recorrente na área de ouvidoria pública. Ela consiste na leitura e análise dos textos e dos anexos às denúncias a fim de se decidir sobre a aptidão das denúncias. A automatização desse tipo de atividade, como sugerido em [de Paiva and Pereira, 2021], torna-se uma alternativa. Porém, em

algumas situações essa automatização não é possível, fazendo com que a análise manual seja imprescindível.

A grande quantidade de informações presentes nas denúncias torna a análise manual muito custosa. Dessa forma, surge a necessidade de métodos de sumarização capazes de resumir os textos das denúncias e de seus anexos a fim de facilitarem a atividade de análise de aptidão de denúncias.

Sendo assim, o objetivo desse trabalho é propor e avaliar duas estratégias de sumarização de denúncias a serem utilizadas na análise de aptidão de denúncias.

O restante desse artigo está dividido da seguinte forma: a Seção 2 traz alguns trabalhos relacionados e a Seção 3 apresenta as metodologias de sumarização propostas. Já as Seções 4 e 5 relatam os experimentos e a análise dos seus resultados, respectivamente. Finalmente, a Seção 6 faz a conclusão do trabalho.

2. Trabalhos Relacionados

A tarefa de sumarização pode ser categorizada em dois métodos: extrativo e abstrativo. A sumarização extrativa seleciona frases do documento original para formar um resumo, enquanto a sumarização abstrativa interpreta o documento original e gera o resumo com outras palavras [Abdel-Salam and Rafea, 2022]. Como é muito difícil para a máquina produzir um resumo compreensível para os humanos, na prática, as abordagens extrativas são mais utilizadas [Ghodratnama et al., 2020].

Abdel-Salam and Rafea (2022) definem 3 abordagens principais para a sumarização de textos: as abordagens estatísticas, as baseadas em grafos e as baseadas em *deep learning*.

As abordagens baseadas em estatísticas são as mais antigas, e começaram a ser desenvolvidas na década de 50. Luhn (1958) sugeriu o uso de informações derivadas do cálculo de frequências das palavras e da sua distribuição no texto para calcular uma medida de significância dessas palavras.

Posteriormente, Edmundson (1969) parte da ideia de Luhn (1958) e propõe a seleção automática de sentenças com maior potencial de transmitir o conteúdo dos textos. Edmundson (1969) sugere ainda a identificação de algumas palavras (*"cue words"*) que sinalizam conteúdos importantes.

Outra abordagem baseada em estatística é a apresentada em [Nenkova and Vanderwende, 2005]. Os autores propõem uma estratégia de busca gulosa que classifica as sentenças de acordo com as frequências, a fim de definir os pesos das probabilidades das palavras e minimizar redundâncias.

Steinberger et al. (2004) descrevem um método genérico de sumarização de texto que utiliza a técnica de análise semântica latente para identificar sentenças semanticamente importantes. Os autores aplicam decomposição de valores singulares sobre a matriz do documento, para fazer a identificação das sentenças.

Com relação as abordagens baseadas em grafos, Mihalcea and Tarau (2004) sugerem a representação do documento como um grafo de sentenças, sendo que, as sentenças representam os nós do grafo e a similaridade entre as sentenças são representadas pelas arestas.

Outra abordagem baseada em grafos é apresentada em [Erkan and Radev, 2004]. Esse trabalho utiliza o conceito de autovetores para auxiliar na criação dos grafos representativos dos textos.

Kosmajac and Kešelj (2019) utilizam o algoritmo TextRank [Mihalcea and Tarau, 2004] para propor uma sumarização extrativa de notícias escritas no idioma sérvio. O processo começa com a concatenação dos textos dos artigos. Depois, divide o texto em frases individuais. Posteriormente encontra a representação vetorial dessas sentenças. Então utiliza-se a semelhança dos cossenos para gerar uma matriz de similaridade. Essa matriz é convertida em um grafo. Por fim, as frases melhor classificadas são selecionadas para formar o resumo.

Nessa mesma linha, Dong et al. (2020) também propõem uma forma de ranqueamento de frases baseada em grafos, a fim de sumarizar textos científicos. O modelo assume uma representação gráfica hierárquica de dois níveis para o documento e busca indicações de assimetrias posicionais para determinar a importância das sentenças.

Segundo os autores, os resultados dos experimentos sugerem que os padrões na estrutura do discurso são um forte sinal para determinar a importância das sentenças em artigos científicos.

Atualmente, o avanço das arquiteturas das redes neurais, e dos modelos de linguagem derivados da arquitetura transformers [Vaswani et al., 2017] estão possibilitando o alcance de excelentes resultados nas tarefas de sumarização textual baseadas em *deep learning*.

Nesse sentido, Miller (2019) utiliza o BERT [Devlin et al., 2019], um modelo de linguagem derivado da arquitetura *transformres*, para encontrar a representação vetorial das sentenças dos textos. Após isso, o autor aplica um processo de clusterização nesses vetores. Por fim, são identificados os centroides de cada um dos *clusters* gerados e as sentenças cujos vetores estão mais próximos desses centroides são as que têm melhores condições de representar o texto em questão.

Gu et al. (2021) tenta identificar frases de qualidade dentro de corpus de texto. Para isso, os autores utilizam o ROBERTA [Liu et al., 2019], outro modelo de linguagem derivado da arquitetura transformer, para identificar frases de qualidade. Os autores propõem a captura de frases de alta qualidade com base na força do relacionamento existente entre as palavras de uma mesma sentença.

Todos esses trabalhos têm em comum o fato dos textos sumarizados serem bem estruturados e escritos corretamente. Porém, os textos das denúncias nem sempre apresentam uma estrutura organizada e bem escrita e nem uma sequência coerente de ideias. Isso acontece por deficiências na escrita ou por problemas no processo de conversão dos diferentes formatos de arquivo (pdfs, apresentações, figuras) para o formato textual ou ainda pela junção (automática) de diferentes arquivos, que apesar de comporem uma mesma denúncia, muitas vezes não tratam de assuntos complementares ou correlatos.

Sendo assim, a principal contribuição desse artigo é a identificação da estratégia de sumarização mais adequada para auxiliar na atividade de análise de aptidão de denúncias.

3. Geração de Resumos

Esse trabalho propõe duas técnicas de sumarização de textos para a geração dos resumos de denúncias, uma das técnicas utiliza uma abordagem estatística e baseia-se em frequência de palavras e a outra utiliza uma abordagem de *deep learning* e emprega o modelo de linguagem BERT [Devlin et al., 2019]. As próximas subseções detalham as estratégias propostas.

3.1. Sumarização pela Frequência de Palavras

O método de sumarização pela frequência de palavras baseia-se em um levantamento estatístico das frases mais relevantes do texto e anexos das denúncias. O método em questão está dividido em 5 passos: pré-processamento dos textos, cálculo da frequência ponderada dos *tokens*, identificação das frases, cálculo de importância das frases e seleção de frases.

A ideia principal dessa metodologia é que palavras que aparecem mais vezes no texto tendem a ser mais importantes. Sendo assim, aquelas frases que possuam tais palavras serão consideradas mais relevantes para o contexto e terão condições de transmitir melhor a ideia que está sendo passado no texto.

A Figura 1 ilustra o processo de geração de resumos pela frequência de palavras, sendo que, os números nos círculos pretos representam a ordem em que as atividades ocorrem.

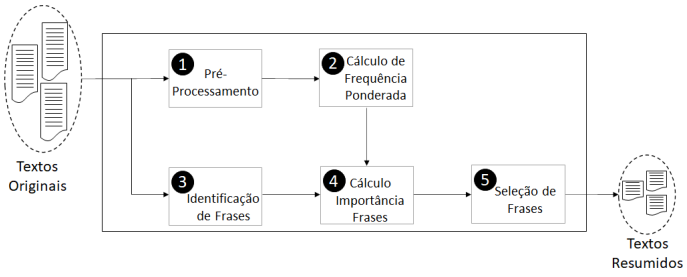

Figura 1. Processo de Sumarização pela Frequência de Palavras

Durante o pré-processamento, é realizado um tratamento no texto de forma que todas as letras sejam passadas para o formato de letra minúscula e que todos os sinais de acentuação sejam retirados. Esse tratamento faz-se necessário para que palavras iguais grafadas de formas diferentes não recebam tratamentos distintos. Nessa fase, também são retiradas as *stopwords* e sinais de pontuação.

O passo seguinte é o cálculo da frequência ponderada de cada uma das palavras que compõem o texto pré-processado. Sendo assim, o texto é *tokenizado* e posteriormente calcula-se a frequência absoluta (quantidade de ocorrências) de cada *token* no texto tratado. Uma vez realizada a contagem dos *tokens*, deve-se identificar o *token* com maior número de ocorrências, sendo que tal *token* receberá o valor 1 e os demais receberão valores de frequências proporcionais. Ou seja, caso o *token* mais frequente tenha aparecido 10 vezes e um outro *token* tenha aparecido 4 vezes, o *token* mais frequente receberá o valor 1 e o outro o valor de 0,4 (4/10).

Na fase de identificação das frases, os textos originais (sem o pré-processamento) são divididos em sentenças. A fase seguinte faz o cálculo de importância de cada uma

das frases dentro do texto. Sendo assim, percorre-se cada um dos *tokens* dessas frases e verifica-se a pontuação de tais *tokens* (de acordo com o valor da frequência ponderada calculado na segunda fase do processo). A importância de cada frase será dada pela soma dos valores de frequência ponderada dos *tokens* que a compõem. Sendo assim, quanto maior for o valor dessa soma, maior será a importância da frase para o texto em questão.

O passo final consiste na ordenação em forma decrescente de importância das frases e na seleção das n frases mais importantes, sendo n um parâmetro de entrada do algoritmo, que indica o número de frases desejadas para compor o resumo.

3.2. Sumarização pelo BERT

A segunda técnica de sumarização proposta utiliza o modelo de linguagem pré-treinado BERT Devlin et al. [2019].

A primeira etapa dessa técnica é dividir o texto em sentenças. Depois aplica-se o modelo BERT a cada uma dessas sentenças a fim de se obter o vetor de *embedding* correspondente a cada uma delas. Esses vetores de *embeddings* são as representações vetoriais das respectivas sentenças, sendo que essas representações preservam o conteúdo semântico. Ou seja, sentenças com significados semelhantes são representadas por vetores próximos no espaço vetorial em questão. Da mesma forma, sentenças com significados muito diferentes são representadas por vetores distantes nesse espaço vetorial.

A partir da representação vetorial de todos as sentenças do texto, calcula-se o vetor médio, a fim de se identificar o vetor que representa a ideia central do texto em questão. Por fim, seleciona-se as n sentenças com menores distâncias para esse vetor médio. Essas sentenças selecionadas são as que melhor representam o texto e por isso são escolhidas para compor o resumo a ser criado. A Figura 2 ilustra esse processo.

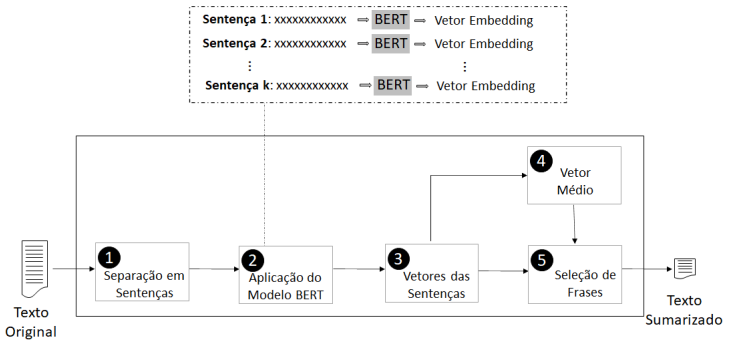

Figura 2. Processo de Sumarização pelo BERT

4. Experimentos

Os processos de sumarização foram avaliados através da análise dos resultados de modelos de classificação textual gerados a partir dos textos originais e dos resumos obtidos pelas metodologias de sumarização propostas. Ou seja, utilizou-se um conjunto de denúncias previamente rotuladas (como aptas ou não aptas), gerou-se resumos dos textos dessas denúncias (pelos dois métodos apesentados) e gerou-se diferentes modelos de classificação a partir dos textos originais e resumidos.

Para a geração dos modelos foram utilizadas duas abordagens distintas: uma que gerava modelos de classificação baseados em *deep learning* (que utilizava uma rede neural com uma camada de BERT, uma camada de *dropout* e uma camada de saída com dois neurônios) e outra que gerava modelos de classificação baseados em técnicas tradicionais de aprendizado de máquinas (que utilizava vetorização TF-IDF[1] e o algoritmo LightGBM).

A fim de evitar conclusões que não reflitam a realidade, os experimentos foram repetidos 30 vezes. Sendo assim, foram gerados 30 modelos distintos, com diferentes divisões de dados de treino e de teste para cada uma das estratégias testadas. Cabe ressaltar que todos os testes foram executados utilizando-se 80% dos dados para o treinamento e 20% para a avaliação. A Figura 3 ilustra a arquitetura dos experimentos.

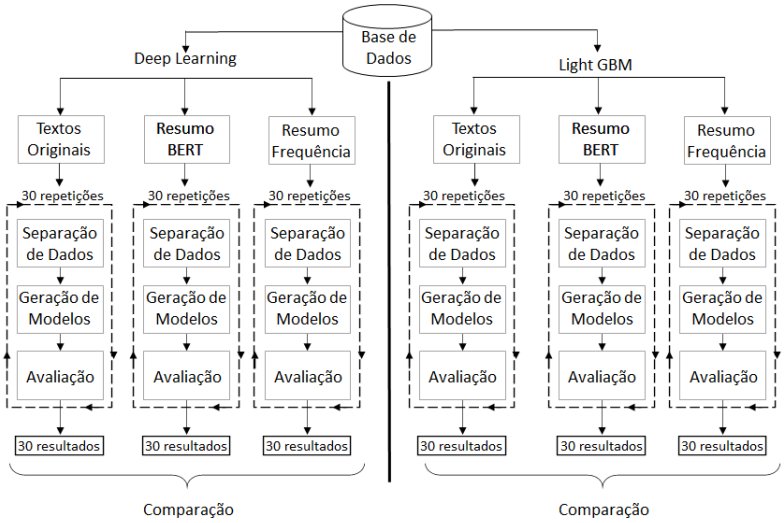

Figura 3. Arquitetura dos Experimentos

Durante os experimentos não se investiu na otimização de hiperparâmetros nem na arquitetura da rede utilizada, pois o objetivo dos testes era verificar o comportamento das soluções propostas em modelos obtidos a partir das mesmas configurações, mas com textos de entrada distintos: textos originais e textos resumidos (obtidos por ambas as técnicas apresentadas).

A base de dados utilizada era composta por uma amostra de 580 denúncias aptas e 580 denúncias consideradas não aptas, selecionadas aleatoriamente. Sendo assim, cada registro dessa base se referia a uma denúncia e era composto pelo texto da denúncia (concatenado com os seus anexos) e um rótulo, que servia para identificar se a denúncia tinha sido considerada como apta ou não.

As denúncias que compunham a base em questão tinham um tamanho médio de 1009 sentenças, enquanto o tamanho mediano dessas denúncias era de 189 sentenças. Essa diferença entre a média e a mediana indica que há poucas denúncias com muitas sentenças, o que acaba deslocando o valor da média muito para a direita.

[1]TF-IDF (Term Frequency – Inverse Document Frequency) é uma técnica de vetorização que indica a importância de uma palavra em um documento em relação a toda coleção de documentos.

Os experimentos foram executados com três tipos de textos: textos originais, textos resumidos pelo modelo BERT e textos resumidos pela frequência. Os resumos foram gerados pela seleção das 10 frases mais relevantes de cada um dos textos originais (de acordo com as metodologias apresentadas).

A opção pelo número de 10 frases se deu pelo fato desse parâmetro ser o que gera o número de *tokens* mais próximo do limite de *tokens* admitido como entrada para o modelo BERT (512 *tokens*).

5. Análise dos Resultados

A métrica utilizada para a avaliação foi a ROC-AUC. Sendo assim, foram utilizados os resultados obtidos pela métrica ROC-AUC durante as 30 execuções de cada um dos experimentos, para as duas estratégias analisadas: modelos obtidos com *deep learning* e modelos obtidos por técnicas tradicionais de aprendizado de máquina.

5.1. Análise dos resultados dos modelos de *Deep Learning*

A Figura 4 apresenta os Box-plots com a consolidação dos resultados das 30 execuções de cada um dos experimentos com os modelos de classificação com *deep learning*.

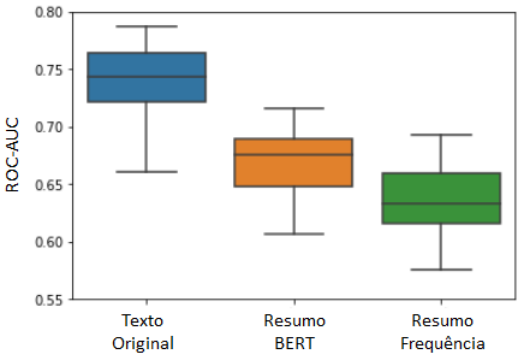

Figura 4. Box-plots com os resultados da Métrica ROC-AUC para os experimentos com *Deep Learning*

Analisando-se a Figura 4, intui-se que o desempenho dos modelos de classificação obtidos a partir dos textos originais foi o melhor e que os modelos obtidos com o resumo pelo BERT obtiveram melhores resultados que os obtidos pelo resumo pela frequência de palavras.

5.1.1. Comparação dos Resultados dos Modelos dos Resumo pelo BERT e do Resumo pela Frequência de palavras

Apesar dos resultados apresentados sugerirem que o modelo obtido com o resumo BERT é melhor do que o modelo obtido pelo resumo pela frequência de palavras, faz-se necessária a realização de testes estatísticos que possam indicar se essas diferenças são estatisticamente relevantes ou se são frutos do acaso.

Sendo assim, optou-se pela realização do teste t para 2 amostras, que tem o objetivo de comparar os valores médios dessas amostras. No entanto, esse tipo de teste parte

do pressuposto de que a distribuição dos valores analisados obedece a uma distribuição normal. Sendo assim, a primeira análise a ser feita é a análise de normalidade dos dados estudados.

Para essa análise de normalidade, realizou-se o teste de Shapiro Wilk. Esse teste considera as seguintes hipóteses nula (H_0) e alternativa (H_1)

$$\begin{cases} \boldsymbol{H_0} : \text{a amostra provém de uma população normal} \\ \boldsymbol{H_1} : \text{a amostra não provém de uma população normal} \end{cases}$$

Todos os testes realizados nesse estudo consideraram o nível de significância de 5% ($\alpha = 0,05$). Sendo assim, caso o valor-p[2] obtido pelo teste estatístico seja menor do que esse nível de significância, pode-se dizer que a hipótese nula (H_0) foi rejeitada e consequentemente a hipótese alternativa (H_1) deve ser considerada como verdadeira. Da mesma forma, caso o valor-p seja maior do que o nível de significância, não é possível rejeitar a hipótese nula, e consequentemente, considera-se ela como sendo verdadeira.

Os testes de Shapiro Wilk aplicados aos resultados dos experimentos com os resumos com o BERT e com o resumo pela frequência de palavras foram de 0,73 e 0,57, respectivamente. Ou seja, como esses valores são maiores que o nível de significância considerado, a hipótese nula não deve ser rejeitada. Logo, pode-se dizer que ambas as amostras provêm de uma população normalmente distribuída.

Uma vez comprovada a normalidade dos dados, outra análise necessária para a realização do teste t é a análise das variâncias das amostras a serem comparadas, pois o teste t varia de acordo com a homogeneidade ou não das variâncias dessas amostras. Para a análise da variância, executou-se o teste F de Levene, que apresenta as seguintes hipóteses nula e alternativa:

$$\begin{cases} \boldsymbol{H_0} : \text{as variâncias são iguais} \\ \boldsymbol{H_1} : \text{as variâncias são diferentes} \end{cases}$$

Nesse teste, obteve-se o valor-p de 0,51, que é maior do que o nível de significância considerado. Logo, a hipótese nula não pode ser rejeitada e consequentemente as variâncias das amostras analisadas podem ser consideradas como homogêneas.

Por fim, após a comprovação de atendimento dos pré-requisitos, aplicou-se o teste t unicaudal para duas amostras, com as seguintes hipóteses formuladas, onde μ indica a média dos resultados obtidos:

$$\begin{cases} \boldsymbol{H_0} : \mu_{\text{Resumo BERT}} = \mu_{\text{Resumo Frequência}} \\ \boldsymbol{H_1} : \mu_{\text{Resumo BERT}} > \mu_{\text{Resumo Frequência}} \end{cases}$$

O teste em questão retornou um valor-p bem próximo de zero. Sendo assim, como o valor-p foi menor do que o nível de significância ($\alpha = 0,05$) a hipótese nula deve ser

[2]O valor-p indica a probabilidade de se observar uma diferença tão grande ou maior do que a que foi observada sob a hipótese nula[Ferreira and Patino, 2015].

rejeitada e consequentemente, assume-se a hipótese alternativa como verdadeira. Logo, pode-se afirmar, com um grau de confiança de 95% (visto que o nível de significância é de 5%) que as médias dos resultados dos modelos obtidos com o resumo com BERT realmente é maior do que a média dos resultados obtidos com o resumo pela frequência de palavras.

5.1.2. Comparação dos Resultados dos Modelos do texto Original com dos Resumo BERT

Para essa análise também foi realizado o teste t. Dessa forma, mais uma vez, fez-se necessária a verificação dos pressupostos do referido teste. Quanto a análise de normalidade, só se analisou a normalidade dos resultados obtidos com o texto original, visto que, a normalidade dos resultados do resumo com BERT já havia sido verificada na subseção anterior.

O teste de Shapiro Wilk retornou um valor-p de 0,31, que é maior do que o nível de significância considerado, o que aponta para a normalidade dos dados.

Para o teste F de Levene, que avalia a homogeneidade de variâncias (no caso entre os resultados obtidos pelos experimentos dos resumos feitos pelo BERT e dos textos originais) obteve-se o valor-p de 0,60, que indica que as variâncias das amostras são homogêneas.

Por fim, aplicou-se o teste t para duas amostras com a seguinte formulação das hipóteses:

$$\begin{cases} \boldsymbol{H_0} : \mu_{\text{Texto Original}} = \mu_{\text{Resumo BERT}} \\ \boldsymbol{H_1} : \mu_{\text{Texto Original}} > \mu_{\text{Resumo BERT}} \end{cases}$$

Para esse teste, obteve-se um valor-p bem próximo de zero, o que nos permite rejeitar a hipótese nula e concluir que a média do resultado obtido pelo texto Original realmente é maior do que a média obtida pelo modelo do Resumo BERT (com nível de confiança de 95%).

5.2. Análise dos resultados dos Modelos Tradicionais

A Figura 5 apresenta os Boxplots dos resultados com modelos tradicionais. Conforme pode ser observado, novamente, os resultados indicam um melhor desempenho dos modelos obtidos pelo texto original, seguidos pelos modelos obtidos pelo resumo BERT e pelo resumo pela frequência de palavras. Porém, para sustentar tais conclusões, foram realizados testes estatísticos.

5.2.1. Comparação dos Resultados dos Modelos dos Resumo pelo BERT e do Resumo pela Frequência de palavras

Para essa comparação também se utilizou o teste t. Desta forma, foram realizados os testes de normalidade e de homogeneidade de variância, citados anteriormente, sendo que

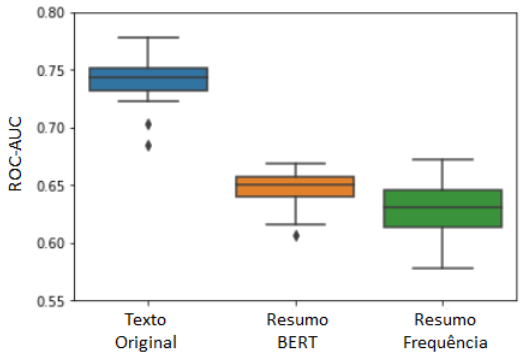

Figura 5. Box-plots com os resultados da Métrica ROC-AUC para os experimentos com modelos tradicionais

as hipóteses nula e alternativa de tais testes foram as mesmas e por esse motivo não foram apresentadas novamente.

Para o teste de Shapiro Wilk obteve-se o valor-p de 0,14 e 0,92 para as distribuições do resumo pelo BERT e do resumo pela frequência de palavras, respectivamente. Logo, pode-se concluir que essas distribuições são normais (visto que esses valores são maiores do que o nível de significância de 0,05).

Com relação a análise de homogeneidade de variância das amostras, realizou-se o teste F de Levene e obteve-se o valor-p de 0,04, que é menor do que o nível de significância considerado ($\alpha = 0,05$), fazendo com que a hipótese nula seja rejeitada e que a hipótese alternativa seja considerada como verdadeira. Ou seja, as variâncias das amostras não são homogêneas.

Por fim aplicou-se o teste t para o caso de amostras com variâncias diferentes, sendo que as hipóteses consideradas foram as seguintes:

$$\begin{cases} \boldsymbol{H_0} : \mu_{\text{Resumo Original}} = \mu_{\text{Resumo BERT}} \\ \boldsymbol{H_1} : \mu_{\text{Resumo Original}} > \mu_{\text{Resumo BERT}} \end{cases}$$

O valor-p para o teste em questão foi de 0,0002. Logo, como esse valor é menor do que o nível de significância considerado, pode-se dizer que a hipótese nula foi rejeitada e assume-se a hipótese alternativa como verdadeira. Ou seja, a média dos resultados obtidos pelos modelos do resumo pelo BERT é maior do que a média dos resultados obtidos pelos modelos do resumo pela frequência de palavras (com um grau de confiança de 95%).

5.2.2. Comparação dos Resultados dos Modelos do texto Original com dos Resumo BERT

Para aplicar o teste t na comparação entre o resultado dos modelos obtidos pelo texto original com os resultados dos modelos obtidos pelo resumo pelo BERT, fez-se a análise de desses resultados, pela aplicação do teste de Shapiro Wilk. Nesse caso, obteve-se o valor-p de 0,02, que é menor do que o nível de significância. Esse resultado indica que a hipótese nula deve ser rejeitada. Logo, os resultados do texto original não proveem de

uma distribuição normal e por isso não se pode aplicar o teste t.

Sendo assim, optou-se pela aplicação de um teste não paramétrico, que possui menos pressupostos com relação aos dados testados. Utilizou-se o teste de Wilcoxon, que em vez de comparar as médias das amostras, compara as suas medianas. Para a realização desse teste, formulou-se as seguintes hipóteses:

$$\begin{cases} \boldsymbol{H_0} : Mediana_{\text{Texto Original}} = Mediana_{\text{Resumo BERT}} \\ \boldsymbol{H_1} : Mediana_{\text{Texto Original}} > Mediana_{\text{Resumo BERT}} \end{cases}$$

O teste retornou um valor-p de bem próximo de zero. Com isso, a hipótese nula foi rejeitada e consequentemente optou-se pela hipótese alternativa, que diz que a mediana dos resultados dos modelos com o texto original é maior do que a mediana dos resultados dos modelos obtidos com o resumo pelo BERT.

5.3. Conclusão da Análise

Tanto o experimento com *deep learning*, quanto o experimento com modelos tradicionais indicaram um desempenho superior para os modelos obtidos com o texto original. Essa indicação já era esperada, uma vez que todo processo de sumarização pressupõe alguma perda de informação em relação ao texto original.

No entanto, também ficou comprovado que o resumo obtido pela metodologia do BERT apresentou melhores resultados que o resumo obtido pela metodologia da frequência de palavras. Isso demonstra que a técnica do resumo pelo BERT é a mais apropriada para auxiliar a análise de aptidão de denúncias.

6. Conclusão

Esse artigo apresentou a proposta de dois métodos de sumarização textual a serem aplicados nos textos de denúncias: um baseado na utilização do modelo BERT e outro baseado em frequência de palavras.

Esses métodos foram avaliados pelos resultados dos modelos de classificação textual de aptidão de denúncias gerados a partir de textos originais e resumidos. Os modelos de classificação foram gerados com a utilização de duas abordagens: uma baseada em *deep learning*, com a utilização do modelo BERT e outra baseada em métodos tradicionais de aprendizado de máquina, utilizando vetorização TF-IDF e o algoritmo Light GBM.

Cada experimento foi realizado 30 vezes, a fim de se minimizar o efeito da aleatoriedade nos resultados obtidos. Por fim, aplicou-se teste estatísticos aos conjuntos de resultados, a fim de se comparar as performances.

Os testes apontaram, com um nível de confiança de 95%, que o desempenho dos modelos gerados pelos textos originais era melhor do que o desempenho dos modelos gerados pelos resumos, e que o desempenho dos modelos gerados pelo resumo do BERT era melhor do que o gerado pelo resumo pela frequência de palavras.

Referências

Abdel-Salam, S. and Rafea, A. (2022). Performance study on extractive text summarization using bert models. *Information*, 13(2):67.

de Paiva, E. and Pereira, F. S. (2021). Extraction and enrichment of features to improve complaint text classification performance. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 338–349. SBC.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Dong, Y., Mircea, A., and Cheung, J. C. (2020). Discourse-aware unsupervised summarization of long scientific documents. *arXiv preprint arXiv:2005.00513*.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Ferreira, J. C. and Patino, C. M. (2015). What does the p value really mean? *Jornal Brasileiro de Pneumologia*, 41(5):485.

Ghodratnama, S., Beheshti, A., Zakershahrak, M., and Sobhanmanesh, F. (2020). Extractive document summarization based on dynamic feature space mapping. *IEEE Access*, 8:139084–139095.

Gu, X., Wang, Z., Bi, Z., Meng, Y., Liu, L., Han, J., and Shang, J. (2021). Ucphrase: Unsupervised context-aware quality phrase tagging. *arXiv preprint arXiv:2105.14078*.

Kosmajac, D. and Kešelj, V. (2019). Automatic text summarization of news articles in serbian language. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–6. IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Steinberger, J., Jezek, K., et al. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4(93-100):8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.