Ferramenta para Classificação de Denúncias: Uma abordagem Baseada em Textos e Dados Estruturados

Eduardo de Paiva[1], **Nelson Ebecken**[1]

[1]COPPE – Universidade Federal do Rio de Janeiro (UFRJ)
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro, Brasília – DF - Brasil

eduardo.paiva@coc.ufrj.br, nelson@ntt.ufrj.br

Abstract. *In Brazil, citizens can report irregularities in the Public Administration. However, these complaints need a prior analysis. This analysis is costly and should consider other information out of the complaints' text. Therefore, the purpose of this research is to develop a model for complaints classifying composed by two other models' combination: one based on structured data, which are obtained by elements extracted from complaints texts and complemented with information from external databases and another obtained directly by the complaints' texts processing.*

Resumo. *No Brasil, os cidadãos podem fazer denúncias de irregularidades na Administração Pública. Porém, para serem apuradas, essas denúncias precisam de uma análise prévia. Essa análise é custosa e considera outras informações que não estão nos textos das denúncias. Sendo assim, o objetivo dessa pesquisa é desenvolver um modelo de classificação de denúncias que é composto pela combinação de outros dois modelos: um baseado em um conjunto de dados estruturados, que são obtidos a partir de elementos extraídos dos textos das denúncias e complementados com informações de bases de dados externas e outro obtido pelo processamento direto dos textos das denúncias.*

1. Introdução

Essa pesquisa tem o objetivo de propor um modelo de classificação de denúncias que é composto pela combinação de outros dois modelos: um baseado em um conjunto de dados estruturados, que são obtidos a partir de elementos extraídos dos textos das denúncias e complementados com informações de bases de dados externas e outro obtido pelo processamento direto dos textos das denúncias.

No Brasil, qualquer cidadão pode fazer denúncias sobre irregularidades que estejam acontecendo na Administração Pública. No entanto, para uma denúncia ser apurada, ela precisa de uma análise prévia. Sendo assim, essa pesquisa se propõe a estudar formas de automatizar essa análise.

A tarefa em questão consiste em um problema de classificação textual, porém, algumas peculiaridades a distinguem dos casos de classificação textual já abordados na literatura. Essa classificação deve levar em consideração outras informações que não estão presentes no conteúdo dos textos. Atualmente, existem vários estudos que tratam sobre classificação textual, no entanto, conforme apresentado na Seção de trabalhos relacionados de [de Paiva and Pereira, 2021], esses estudos não consideram informações externas ao texto para a realização da classificação.

Para o caso em questão, um processo automatizado deve identificar e extrair certas informações do texto das denúncias. Feldman et al. (2007) apresenta quatro tipos básicos de elementos que podem ser extraídos de textos: entidades, atributos, fatos e eventos. Dessa forma, esse trabalho busca por esses elementos nos textos das denúncias e depois tenta correlacioná-los com outras informações em bases de dados externas. Além disso, essa pesquisa também estuda diferentes formas de tratar os textos originais das denúncias.

O restante dessa proposta de tese está dividido da seguinte forma: a Seção 2 define o problema a ser tratado. As Seções 3 e 4 descrevem a proposta de pesquisa e o projeto de avaliação, respectivamente. Finalmente, a Seção 5 faz a conclusão do trabalho.

2. Definição do Problema

No âmbito federal, a Ouvidoria-Geral da União (OGU) é responsável por receber e tratar as denúncias referentes a agentes públicos, órgãos e entidades do Poder Executivo Federal.

No entanto, para saber se tais denúncias possuem informações consistentes para serem apuradas, faz-se necessária a identificação, validação e análise das situações narradas. Essa avaliação exige um grande esforço e dispêndio de tempo por parte de servidores da OGU.

Durante a avaliação, o servidor deve ler o texto da denúncia e acessar e analisar cada arquivo anexo à denúncia. Esses arquivos anexos podem estar em diferentes formatos (planilhas, figuras, apresentações, arquivos textos e etc...).

Após a análise dos textos das denúncias, o servidor deve verificar as informações relatadas nesses documentos em bases de dados e sistemas corporativos. A partir dessas análises, ele conclui sobre a aptidão ou não das denúncias.

Sendo assim, o problema que essa pesquisa se propõe a tratar é: como criar um modelo capaz de predizer, a partir dos textos das denúncias e de seus respectivos anexos se essas denúncias devem ser consideradas como aptas ou não.

Logo, o problema em questão é caracterizado como um problema de classificação textual. No entanto, esse problema de classificação tem alguns dificultadores. O primeiro é que a classificação deve levar em consideração não apenas o conteúdo das denúncias, mas também o conteúdo dos arquivos anexos, que podem estar em diferentes formatos.

O outro dificultador reside no fato de que essa classificação não pode se limitar ao conteúdo dos textos. Ela tem que ser capaz de extrair certas informações dos textos e correlacioná-las com outras bases de dados, a fim de fornecer mais parâmetros para a tomada de decisão.

Mais uma característica dos textos das denúncias que precisa ser considerada é que eles costumam ser muito grandes. Ou seja, a solução em questão deve ser capaz de tratar textos longos.

Cabe ressaltar que em muitas situações os textos possuem informações redundantes ou irrelevantes para a análise em questão, sendo necessária a identificação das partes mais importantes a serem consideradas pelo processo.

Sendo assim, formulou-se três questões de pesquisa a serem respondidas durante

os estudos:

- **Questão 1:** Como considerar informações externas aos textos no processo de classificação das denúncias?
- **Questão 2:** Como extrair as principais informações dos textos das denúncias a fim de utilizá-las no processo de classificação textual?
- **Questão 3:** Como desenvolver uma solução de classificação textual capaz de tratar textos longos?

A justificativa pela escolha de cada uma dessas questões de pesquisa é apresentada na Seção 3, mais especificamente nas Subseções que tratam de cada uma dessas questões.

3. Projeto de Pesquisa

Essa pesquisa pretende desenvolver um classificador textual que é obtido a partir da combinação de dois modelos de classificação: um baseado em dados estruturados e outro baseado nos dados textuais originais

No entanto, a obtenção desses modelos classificação depende das respostas das três questões de pesquisa apresentadas na Seção 2. Sendo assim, a resposta da questão 1 possibilita a obtenção do modelo baseado nos dados estruturados e as respostas das questões 2 e 3 são úteis para o desenvolvimento do modelo baseado nos dados textuais originais.

Logo, esse projeto de pesquisa tem o objetivo de responder a essas questões, sendo que a parte da pesquisa referente a questão 1 já foi realizada, e a pesquisa referente as demais questões ainda está sendo feita.

3.1. Questão 1: Como considerar informações externas aos textos no processo de classificação das denúncias?

Em muitas situações, a classificação textual carece de informações que não estão narradas diretamente nos textos, ou dependem de formas alternativas para a representação desses textos.

Nesse sentido, alguns trabalhos tentam contornar tais problemas. Wu et al. (2018) propõem a criação de um dicionário de gírias que é consultado durante a classificação.

Karthikeyan et al. (2019) propõem a classificação de textos da internet utilizando apenas partes do conteúdo desses textos. Os autores extraem os documentos da web e recuperam os conteúdos relatados com base em *queries*, agregações e transformações de dados, a fim de obter uma representação estruturada do dado anteriormente desestruturado.

Coussement and Van den Poel (2008) sugerem um modelo de classificação que utiliza como variáveis informações sobre o estilo de escrita dos textos.

Esses trabalhos propõem formas alternativas para a representação dos textos a serem classificados. No entanto, essa representação fica limitada a conteúdos extraídos dos próprios textos, ou a dicionários fixos, não considerando informações derivadas ou correlacionadas ao texto analisado. Por tal razão, optou-se por essa questão de pesquisa.

Cabe ressaltar que essa questão de pesquisa já foi respondida e é apresentada de forma mais detalhada em [de Paiva and Pereira, 2021]. O método proposto extrai os 4

tipos de elementos citados por Feldman et al. (2007) : entidades, atributos, fatos e eventos. No entanto, ao invés de realizar essa extração utilizando apenas os textos originais, propõe-se a utilização de fontes externas (57 bases de dados), a fim de se identificar novos elementos relacionados aos que já foram extraídos dos textos originais. Sendo assim, tem-se dois tipos de elementos: os de 1º nível (extraídos diretamente dos textos) e os de 2º nível (oriundos de fontes de dados externas). A metodologia proposta é composta de 5 fases e é ilustrada na Figura 1. Essa metodologia recebe como entrada as denúncias e seus respectivos anexos e fornece como saída um conjunto de dados estruturados capaz de representar cada uma das denúncias.

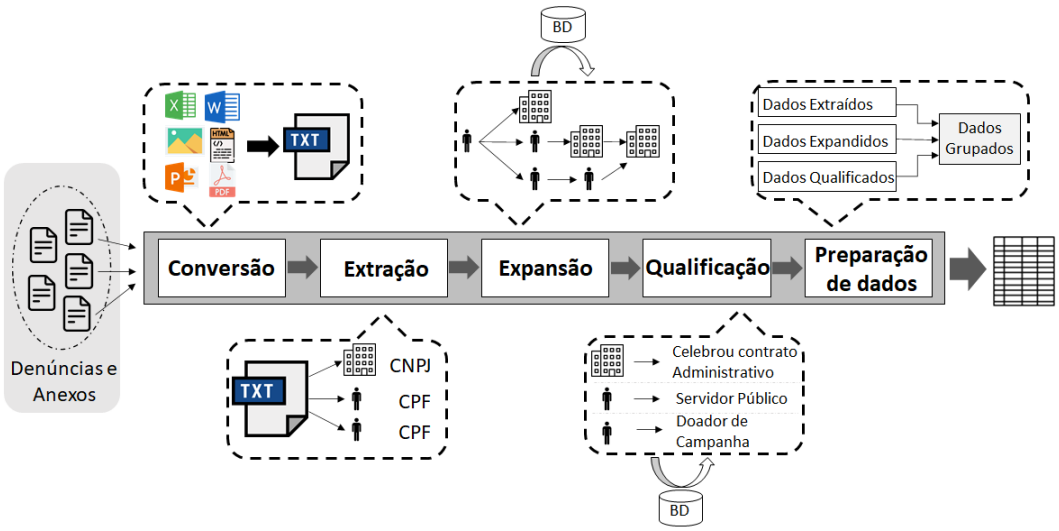

Figura 1. Processo de extração e enriquecimento de variáveis

A primeira fase do processo é a de **Conversão**. A principal função dessa fase é acessar os arquivos anexos e transformá-los em um formato apropriado para a leitura de máquinas.

Esses arquivos anexos podem vir em diferentes formatos (fotos, planilhas, arquivos pdf scaneados como figuras, apresentações etc.). Essa diversidade de formatos geralmente não está preparada para a leitura automatizada de máquinas, o que torna inviável a sua utilização em um processo de descoberta de conhecimento.

Sendo assim, durante essa fase, todo o conteúdo textual desses arquivos é transformado no formato de texto plano, para que possam ser utilizados nas fases posteriores do processamento.

A segunda fase faz a **Extração** de informações dos textos. Essa metodologia faz a identificação e extração de um conjunto de elementos considerados relevantes para a atividade de tratamento de denúncias. Sendo assim, alguns exemplos de elementos extraídos são: nomes de pessoas, CPFs, CNPJs, números de contratos, números de convênios, valores monetários e etc.

A fase de **Expansão** utiliza as entidades identificadas na fase anterior, e tenta encontrar novas informações a respeito delas em outras bases de dados. Essa busca tem o objetivo de validar a existência das entidades identificadas, bem como descobrir novos elementos que tenham vínculos com as entidades identificadas anteriormente.

Logo, para um determinado CNPJ, identificado no texto da denúncia, a expansão realiza duas atividades. Primeiro, ela verifica, em bases de dados institucionais, se esse CNPJ realmente é um CNPJ válido. Posteriormente, são buscados outros elementos derivados desse CNPJ. Por exemplo, identifica-se todas as pessoas que constam como sócias desse CNPJ. Cabe ressaltar que cada tipo de entidade passa por um processo de expansão específico.

A quarta fase do processo tem o objetivo de fazer a **Qualificação** das entidades identificadas nas fases anteriores. Sendo assim, para um determinado CPF, é verificado se ele pertence a um servidor público, se é beneficiário de algum programa social e etc.

Dessa forma, todas as entidades identificadas passam por esse processo de qualificação, sendo que, cada tipo de entidade possui um conjunto específico de qualificadores que são verificados.

A última fase é a de **Preparação dos Dados**. Durante essa fase, agrega-se todas as informações obtidas nas fases anteriores, a fim de se criar um conjunto de dados estruturados que possa ser utilizado no treinamento do modelo.

Sendo assim, cada denúncia passa a ser representada por um conjunto de dados estruturados e pelos textos originais das denúncias (texto principal e textos dos anexos).

3.2. Questão 2: Como extrair as principais informações dos textos das denúncias a fim de utilizá-las no processo de classificação textual?

Atualmente, o avanço das arquiteturas das redes neurais, e dos modelos de linguagem derivados da arquitetura *transformer* [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 *transformer*, para gerar resumos de textos a partir da representação vetorial das sentenças desses textos.

Gu et al. (2021) tentam 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*. Os autores propõem a captura de frases de alta qualidade com base na força do relacionamento existente entre as palavras de uma mesma sentença.

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.

Esses fatores motivaram a escolha dessa questão de pesquisa para compor o trabalho em questão.

Dessa forma, a sumarização a ser desenvolvida será inspirada em [Miller, 2019]. A ideia desse trabalho é separar o texto em sentenças e aplicar o modelo BERT em cada uma das sentenças. Após isso, utiliza-se os vetores de saída (que representam cada uma dessas

sentenças) e aplica-se um processo de clusterização utilizando o algoritmo K-means. O resumo será formado por amostras de cada um dos *clusters*. Também pretende-se testar outras variantes dessa estratégia.

3.3. Questão 3: Como desenvolver uma solução de classificação textual capaz de tratar textos longos?

O estado da arte na área de processamento de linguagem natural aponta para a utilização de modelos de linguagem pré treinados. Esses modelos possibilitam a transferência de aprendizado, potencializando os resultados obtidos pelas aplicações.

No entanto, a única versão de modelo de linguagem atualmente disponível para a língua portuguesa é o apresentado em [Souza et al., 2020], que é uma versão do modelo original do BERT [Devlin et al., 2019]. Porém, o BERT apresenta como limitação o fato de só trabalhar com textos até o limite de 512 tokens, que é um limite muito abaixo do tamanho médio das denúncias.

Já existem outros modelos de linguagem que contornam essa limitação do número de tokens, como por exemplo [Beltagy et al., 2020] e [Zaheer et al., 2020]. No entanto, esses modelos não possuem versões para o idioma português.

Sendo assim, optou-se por essa questão de pesquisa a fim de tornar possível a utilização do modelo de linguagem BERT para textos maiores que o limite de 512 tokens.

Alguns autores já pesquisaram sobre possíveis soluções para esse tipo de limitação do BERT. Nesse sentido, Sun et al. (2019) propõem a divisão do texto de entrada em grupos menores de tokens. Dessa forma, torna-se possível apresentar ao modelo cada um desses grupos separadamente. Assim, obtém-se a representação de cada um dos grupos e em seguida combina-se essas representações (pela aplicação de uma função de agregação) a fim de se obter uma representação final do texto completo.

Ainda nessa linha, Pappagari et al. (2019) segue essa mesma ideia de divisão dos tokens. Porém, em vez de se utilizar uma função de agregação para juntar as diversas saídas do modelo BERT, utiliza-se uma rede LSTM, sendo que, as entradas para essa rede são as saídas do modelo BERT. Dessa forma, a sequência de entrada da rede LSTM é a mesma sequência posicional das saídas do modelo BERT.

Nessa parte da pesquisa, pretende-se testar ambas as estratégias apresentadas. Além disso, pretende-se testar uma variante da abordagem proposta por [Pappagari et al., 2019], sendo que, nesse teste, ao invés de se utilizar uma rede LSTM, utilizar-se-ia uma rede CNN.

4. Projeto de Avaliação

A avaliação será dividida em 4 partes: uma referente a cada uma das questões de pesquisa e outra relativa a solução conjunta do problema. No entanto, todas as partes seguirão os mesmos critérios. Serão realizadas avaliações quantitativas cujo objetivo é verificar o desempenho dos modelos gerados em diferentes situações.

A métrica a ser utilizada será a área sob a curva ROC, que é um valor que pode variar entre 0 e 1 e quanto maior ela for, melhor é a qualidade do modelo.

Os experimentos analisados serão repetidos várias vezes (no mínimo 30 vezes), para gerar um conjunto de resultados que possa ser avaliado pela aplicação de testes

estatísticos. O objetivo dessa prática é minimizar os efeitos oriundos da aleatoriedade. Ainda nesse intuito de minimizar os efeitos da aleatoriedade, todos os experimentos serão executados com validação cruzada.

A Avaliação da questão de pesquisa 1 já foi realizada, uma vez que essa parte da pesquisa já foi desenvolvida. Essa avaliação é descrita em [de Paiva and Pereira, 2021]. Nesse caso, repetiu-se os experimentos 1.000 vezes e a partir dos dados obtidos realizou-se testes de hipótese comparando-se os resultados obtidos com a aplicação da técnica de extração de *features* proposta e sem a aplicação da referida técnica. Os resultados apontaram que a utilização da técnica em questão trouxe ganhos para o processo de classificação.

Para a avaliação da Questão de Pesquisa 2, pretende-se comparar os resultados da classificação textual com e sem a aplicação da técnica de sumarização textual proposta, a fim de verificar se há ganho nos resultados da classificação textual com a aplicação da técnica de sumarização.

Já na avaliação da questão de Pesquisa 3, serão analisadas as diferentes soluções propostas para a classificação de textos longos, a fim de se identificar aquela que fornece melhores resultados.

Por fim, a avaliação da solução conjunta será feita pela utilização de diferentes pesos de ponderação entre o modelo oriundo dos dados estruturados e o modelo oriundo dos dados textuais, a fim de se verificar a combinação que propicia melhores resultados.

5. Conclusão

Este trabalho apresentou os principais passos da pesquisa que está sendo realizada. Essa pesquisa pretende fornecer contribuições acadêmicas e sociais. Quanto a contribuição social, já foi desenvolvida e está em produção uma versão do classificador de denúncias.

Essa pesquisa também já produziu um artigo publicado na revista IEEE Latin America Trasactions [Paiva et al., 2021] e um artigo apresentado no XVIII ENIAC [de Paiva and Pereira, 2021]. Pretende-se publicar ainda dois outros artigos científicos, referentes as questões de pesquisa 2 e 3.

A referida pesquisa pretende apresentar as seguintes contribuições acadêmicas: a proposta de uma técnica de extração e enriquecimento de features em conjuntos de dados textuais; a proposta de um modelo de classificação textual que considera dados estruturados e dados textuais; a proposta de uma arquitetura de rede capaz de receber textos de tamanhos grandes, processá-los utilizando o modelo BERT e fazer a classificação textual e a proposta de uma metodologia de sumarização de textos.

Referências

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Coussement, K. and Van den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4):870–882.

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.

Feldman, R., Sanger, J., et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

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*.

Karthikeyan, T., Sekaran, K., D., R., V., V. K., and M, B. J. (2019). Personalized content extraction and text classification using effective web scraping techniques. *Int. J. Web Portals*, 11(2):41–52.

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.

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

Paiva, E., Paim, A., and Ebecken, N. (2021). Convolutional neural networks and long short-term memory networks for textual classification of information access requests. *IEEE Latin America Transactions*, 19(5):826–833.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11856 LNAI(2):194–206.

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.

Wu, L., Morstatter, F., and Liu, H. (2018). Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Lang. Resour. Evaluation*, 52(3):839–852.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.