Extraction and enrichment of features to improve complaint text classification performance

Eduardo de Paiva[1], **Fernando Sola Pereira**[1]

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

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

Abstract. *In Brazil, citizens can report irregularities in the Public Administration. The classification of these complaints needs information that is not in their texts. this article aims is to propose a methodology for extracting and enriching information identified in complaints texts. This methodology provides as output a set of structured data capable of characterizing the complaints. To validate the proposal, a case study was done. The study showed that structured data use enabled an improvement in the performance of the complaint's classification.*

Resumo. *No Brasil, os cidadãos podem fazer denúncias de irregularidades na Administração Pública. A classificação dessas denúncias necessita de informações que não estão nos seus textos. O objetivo desse artigo é propor uma metodologia para a extração e enriquecimentos de informações identificadas nos textos das denúncias. Essa metodologia fornece como saída um conjunto de dados estruturados capazes de caracterizar as denúncias. Para validar a proposta, foi realizado um estudo de caso. O estudo demonstrou que a utilização dos dados estruturados possibilitou uma melhora no desempenho da classificação das denúncias.*

1. Introdução

Este artigo propõe uma metodologia que faz a extração de variáveis de textos de denúncias e posteriormente correlaciona essas variáveis com outras bases de dados a fim de obter novas informações, para serem utilizadas na classificação textual das denúncias.

O ordenamento jurídico do Brasil permite que qualquer cidadão possa fazer denúncias sobre irregularidades que estejam ocorrendo na Administração Pública.

Essa possibilidade é uma modalidade de controle social. Nesse tipo de controle, o próprio cidadão tem a possibilidade de verificar a regularidade da atuação da Administração, a fim de impedir a prática de atos ilegítimos, lesivos aos indivíduos ou a coletividade ou que possibilitem a reparação dos danos decorrentes da prática de tais atos [Alexandrino and Paulo, 2006].

Visando operacionalizar essa modalidade de controle, foram instituídas as ouvidorias públicas. As ouvidorias públicas são canais que facilitam a interação entre os cidadãos e a Administração Pública, oferecendo assim uma forma para que o cidadão possa exercer o controle social.

No âmbito federal, esse papel é desempenhado pela Ouvidoria-Geral da União (OGU), ligada à Controladoria-Geral da União (CGU). Sendo assim, a OGU recebe uma

série de denúncias que precisam ser tratadas, analisadas e apuradas. Nesse sentido, uma das primeiras atividades referentes ao tratamento das denúncias é a análise de aptidão. Nessa etapa examina-se todo o material referente a uma determinada denúncia (textos das denúncias e anexos auxiliares) a fim de verificar se tal denúncia reúne os requisitos mínimos para o prosseguimento do rito apuratório.

O volume de material a ser analisado por denúncia, aliado à grande quantidade de denúncias que são recebidas, dificulta uma resposta tempestiva para essa análise de aptidão. Sendo assim, faz-se necessário um mecanismo automatizado para a realização de tais análises. Dessa forma, o desenvolvimento de técnicas de processamento de linguagem natural capazes de realizar a classificação de uma denúncia como apta ou não se torna um problema em aberto.

No entanto, para realizar essa análise de aptidão, deve-se validar as informações narradas nos textos e complementá-las com outros conhecimentos que não estão contidas nos limites dos textos das denúncias. Ou seja, deve-se identificar informações específicas nos textos e correlacioná-las com dados externos.

Logo, 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.

Sendo assim, o objetivo desse artigo é propor uma metodologia para a extração e enriquecimento de informações de textos de denúncias, a fim de utilizá-las juntamente com os textos originais para realizar a classificação automatizada de aptidão de denúncias.

Para isso, formulou-se a seguinte hipótese: se forem empregadas técnicas de extração e enriquecimento de variáveis em textos de denúncias, então, os resultados da classificação textual dessas denúncias podem melhorar.

O restante desse artigo está dividido da seguinte forma: a Seção 2 apresenta uma revisão da literatura. Já as Seções 3 e 4 descrevem a proposta e um estudo de caso para validar essa proposta, respectivamente. Finalmente, a Seção 5 faz a conclusão do trabalho.

2. Trabalhos Relacionados

A área de processamento de linguagem natural vem experimentando grandes avanços. Tanto as formas de representação textual quanto a maneira de processar os textos têm evoluído.

Quanto a representação textual, foram desenvolvidas formas capazes não só de representar os textos em formatos numéricos, mas também de capturar informações semânticas e sintáticas das palavras. Dentre essas representações, pode-se citar o word2vec [Mikolov et al., 2013a] e [Mikolov et al., 2013b], glove [Pennington et al., 2014], fasttext [Bojanowski et al., 2017] e o wang2vec [Ling et al., 2015]. As chamadas word embeddings contextualizadas, como por exemplo o BERT [Devlin et al., 2019], conseguem ainda representar diferentes significados de uma determinada palavra de acordo com o contexto em que ela aparece.

Quanto ao processamento dos textos, as arquiteturas de redes neurais profundas

têm proporcionado grandes avanços. As Recurrents Neural Network-RNN [Elman, 1990], mais especificamente as Long Short-Term Memory-LSTM [Hochreiter and Schmidhuber, 1997] e Gated Recurrent Unit – GRU [Chung et al., 2014], e as Convolutional Neural Network-CNN [Lecun et al., 1998] têm sido empregadas com grande sucesso em diversas áreas de processamento de textos. Mas recentemente, as arquiteturas ELMO [Peters et al., 2018], GPT [Radfort et al., 2018] e Transformers [Vaswani et al., 2017] têm obtido excelentes resultados.

Todos esses avanços foram de extrema importância para o aprimoramento do processamento de linguagem natural e consequentemente para a tarefa de classificação de textos. No entanto, mesmo com todos esses avanços, muitas das vezes 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. Esse dicionário é utilizado durante a atividade de classificação de sentimentos de mensagens postadas em mídias sociais. Dessa forma, ao analisar um determinado texto, pode-se consultar o dicionário e encontrar a palavra correspondente a uma determinada gíria citada na mensagem. Esse procedimento facilita a tarefa de classificação de sentimentos.

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.

Após isso, os autores utilizam uma técnica de eliminação recursiva de variáveis com o objetivo de selecionar o melhor sub conjunto de variáveis candidatas. Por fim, o modelo de classificação é gerado com algoritmos da aprendizagem de máquina tradicionais para categorizar as páginas da web.

Li (2019) apresenta um estudo para a área jurídica cujo objetivo é encontrar correspondência entre uma previsão legal e a descrição de um evento escrito na língua inglesa. Para isso, o autor utiliza um método de classificação de textos legais baseado em palavras características.

As palavras características dos documentos são calculadas pelo TF-IDF, porém, o autor propõe a introdução de fatores de correção ao TF-IDF baseados na estatística chi-quadrada e na posição das palavras.

Como as previsões legais podem ser extraídas dos documentos de julgamento com precisão, torna-se possível estabelecer uma relação entre a previsão legal e as palavras características.

Nessa mesma linha de utilizar modificações do TF-IDF, Liu et al. (2018) sugerem a realização da classificação textual utilizando uma representação vetorial de palavras que combina o word2vec com o TF-IDF. O artigo propõe que a representação da palavra seja dada pela multiplicação do peso da palavra, obtido com o TF-IDF, pelo vetor de representação word2vec dessa mesma palavra. Sendo assim, cada texto passa a ser representado pelo acúmulo de todos os vetores de palavras e tal representação pode ser

utilizada na classificação dos textos.

Coussement and Van den Poel (2008) sugerem um modelo de classificação que utiliza como variáveis informações sobre o estilo de escrita dos textos. Sendo assim, são utilizadas como variáveis a quantidade de verbos, pronomes, palavras únicas, palavras de negação, palavras afirmativas, artigos, preposições entre outras informações derivadas do texto. Os autores alegam que a utilização destes elementos de estilo linguístico, juntamente com vetores de palavras, pode aumentar o desempenho do classificador.

Todos esses trabalhos propõem formas alternativas para a representação dos textos a serem classificados. No entanto, essa representação sempre fica limitada a conteúdos extraídos dos próprios textos, sem buscar informações derivadas ou correlacionadas.

Dessa forma, nossa principal contribuição é a proposta de uma metodologia que possibilita a representação dos textos com elementos do próprio texto e com outros elementos correlacionados, extraídos de fontes de dados externas. Sendo assim, essa metodologia propicia a obtenção de um conjunto de dados estruturados que pode ser utilizado para melhorar o desempenho da classificação textual.

3. Proposta

Para o entendimento completo dos textos, muitas vezes é necessário o conhecimento de outras informações que não estão presentes nos textos. Da mesma forma, em muitas situações, o processamento de linguagem natural carece de informações que não estão nos textos analisados. Tal situação ocorre na classificação de textos de denúncias.

Essa seção apresenta uma metodologia que faz a extração de variáveis de textos, e enriquece essas variáveis com informações oriundas de diversas bases de dados.

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, propomos a utilização de fontes externas (57 bancos de dados), a fim de identificarmos novos elementos relacionados aos elementos extraídos. Sendo assim, tem-se dois tipos de elementos: os de 1º nível (extraídos diretamente dos textos) e os de 2º nível (oriundos das fontes de dados externas). A metodologia proposta é composta de 5 fases e é ilustrada na Figura 1. As próximas subseções descrevem o papel de cada uma dessas fases.

3.1. Conversão

A primeira fase do processo é a de Conversão. A principal função dessa fase é ler os arquivos anexos (das denúncias) e extrair as informações desses arquivos.

Os 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.

Dessa forma, os anexos são transformados em um formato textual, para que possam ser utilizados nas fases posteriores do processamento. Essa transformação é feita pelo Apache Tika[1], uma solução que recebe diferentes formatos de arquivos e extrai os seus conteúdos textuais.

[1]https://tika.apache.org/

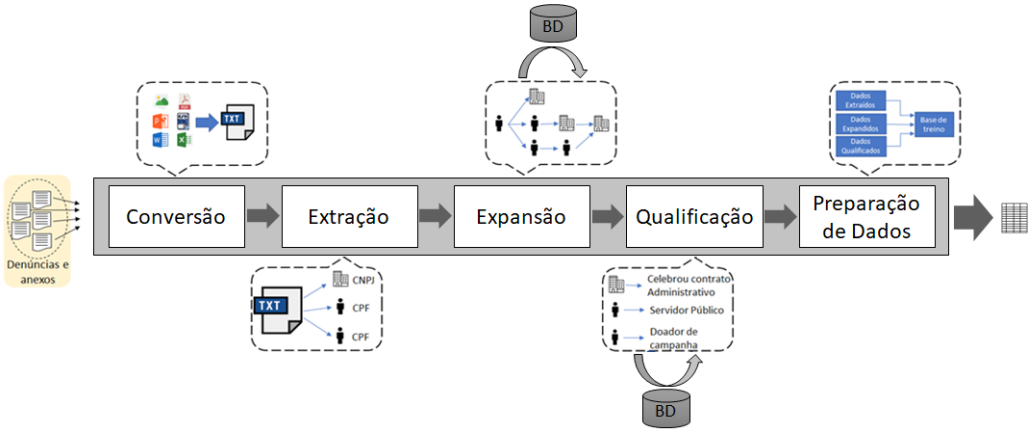

Figura 1. Processo de extração e enriquecimento de variáveis

Após essa atividade, os textos extraídos dos anexos são concatenados aos textos originais das denúncias, a fim de que esses sejam tratados de forma única.

3.2. Extração

Nessa fase é realizada a extração de informações dos textos das denúncias. 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 e empresas
- CPFs
- CNPJs
- Números de NIS[2]
- Números de contratos
- Números de convênios
- Valores monetários
- Palavras fortes

O reconhecimento de nomes é feito conforme a proposta apresentada em [Souza et al., 2019], nesse trabalho os autores fazem o reconhecimento de entidades nomeadas em português empregando o modelo BERT [Devlin et al., 2019], sendo que, nesse caso, utiliza-se uma versão em português do modelo BERT [Souza et al., 2020]. Já o reconhecimento de CPFs, CNPJs, valores monetários, números de NIS, contratos e convênios é realizado pela utilização de expressões regulares. A opção pelo uso de expressões regulares para esse tipo de identificação se deu pelo fato desses elementos apresentarem formatos numéricos bem característicos, de fácil identificação nos textos, tornando o desempenho da identificação melhor do que o obtido por aprendizado de máquina.

Os elementos considerados "Palavras fortes" são palavras, ou expressões, consideradas relevantes no contexto de análise de denúncias (por exemplo "fraude", "corrupção", "super faturamento" e etc). A busca desses tipos de elementos é feita diretamente nos

[2]O NIS é o Número de Identificação Social atribuído pela Caixa Econômica Federal para identificar pessoas cadastradas em programas sociais do governo.

textos, sendo que a definição dessas palavras e expressões foi feita por especialistas da área de negócio.

Uma vez identificados, todos esses elementos são armazenados em um banco de dados para serem utilizadas nas outras fases do processamento.

3.3. Expansão

A 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.

Sendo assim, 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. Da mesma forma, para um determinado número de contrato, verifica-se se esse contrato é valido. Posteriormente, são levantadas as partes envolvidas no contrato (contratante e contratado), o objeto do contrato, prazo de vigência do contrato e etc.

Esse procedimento é executado para todas as entidades identificadas nos textos. Dessa forma, passa-se a ter 2 tipos de elementos: os de primeiro nível (identificados diretamente nos textos) e os de segundo nível (derivados dos anteriores).

3.4. Qualificação

A fase de qualificação 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. Para um determinado CNPJ, é verificado se esse é parte de algum contrato ou convênio, se está cadastrado no CEPIM[3] 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.

3.5. Preparação dos Dados

Durante a preparação dos dados, 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.

Dessa forma, cada tipo de informação levantada nas fases anteriores passa a ser representada como uma coluna e cada denúncia como uma linha. Sendo assim, o valor referente a cada um dos atributos para cada denúncia é dado pela aplicação de alguma função de agregação, que varia de acordo com o tipo de atributo (por exemplo: contagem, soma, média, etc..)

[3]Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM): relação de entidades privadas sem fins lucrativos que estão impedidas de celebrar convênios, contratos de repasse ou termos de parceria com a Administração Pública Federal.

Logo, cada denúncia passa a ser representada por um conjunto de dados estruturados, que foram obtidos nas fases anteriores do processamento, e pelos textos originais das denúncias (texto principal e textos dos anexos).

4. Estudo de Caso

A fim de validar a proposta apresentada e testar a hipótese sugerida, foi realizado um estudo de caso[4].

A intenção desse estudo de caso é executar todos os procedimentos de extração e enriquecimento de variáveis, apresentados na Seção 3, e gerar dois modelos de classificação de denúncias: um obtido pelo processamento dos textos originais das denúncias e outro gerado a partir dos dados estruturados obtidos pela aplicação da metodologia proposta. Por fim, gera-se um novo modelo formado pela combinação dos dois modelos anteriores a fim de verificar se houve melhora no desempenho da classificação com a utilização dos dados estruturados.

A ideia desse modelo combinado se apoia na utilização de modelos ensembles. Domingos (2012) explica que em vez de selecionar um único modelo, é melhor combinar vários modelos, a fim de produzir resultados melhores. Segundo Domingos (2012) , a melhor forma de aumentar a acurácia de um modelo é juntar modelos diferentes a fim de criar um modelo final mais preciso.

As próximas subseções descrevem os passos seguidos durante a execução do estudo de caso.

4.1. Base de Dados

O estudo de caso foi desenvolvido com um conjunto 4051 denúncias (correspondentes ao período de janeiro de 2020 a maio de 2021) que foram direcionadas para a Ouvidoria Geral da União. Essas denúncias eram compostas pelos textos originais, e por seus arquivos anexos. Cada denúncia também possuía um rótulo, informado por especialistas da Ouvidoria, que dizia se a referida denúncia deveria ser considerada como Apta ou Não Apta. Cabe ressaltar que a base de dados em questão era desbalanceada, sendo que, cerca de 30% das denúncias eram consideradas como aptas e 70% como não aptas.

4.2. Métricas de Avaliação

Para direcionar a geração e avaliação dos modelos, foram utilizadas duas métricas: área sob a curva ROC (ROC-AUC) e Área sob a curva Precision-Recall (AUPRC).

A opção por tais métricas se deu pelo fato delas não necessitarem de pontos de corte, ou seja, elas avaliam a probabilidade de um exemplo pertencer a uma determinada classe, sem especificar um determinado ponto de corte. Dessa forma, pode-se fazer a otimização do modelo (sem se preocupar com esse ponto de corte) e ao final escolhe-se o ponto de corte que maximize algum critério de avaliação desejado.

A métrica ROC-AUC indica a chance de um exemplo positivo ter um score de previsão maior do que um exemplo negativo. Essa métrica é indicada para os casos em

[4]O código fonte utilizado no estudo de caso, bem como uma cópia da base de dados descaracterizada encontram-se disponível no seguinte link: https://github.com/fernandosola/eniac2021-article

que se quer garantir que os exemplos positivos estejam ranqueados acima dos exemplos da classe negativa.

Quanto maior o valor dessa métrica, melhor é a performance do modelo, sendo que, o seu valor máximo é 1 e o mínimo deve ser 0,5. Ou seja, caso o valor dessa métrica dê abaixo de 0,5, o modelo está fazendo previsões piores do que se essas previsões fossem feitas de forma aleatória.

A métrica AUPRC faz uma média ponderada da curva de precisão e de *recall*. Essa métrica mostra a comparação entre a precisão e o *recall* para diferentes limites. Uma área alta sob a curva representa alto *recall* e alta precisão. Sendo assim, quanto mais alto o valor da AUPRC, melhor será o modelo.

A métrica ROC-AUC foi utilizada como métrica principal e a métrica AUPRC como métrica secundária.

4.3. Condições de Execução

Os experimentos foram executados com validação cruzada, utilizando-se sempre 5 *folds*, sendo que em cada iteração sempre eram destinados 20% dos dados para treino e 80% para validação.

4.4. Modelo Textual

Para a seleção do algoritmo de geração do modelo textual foram testados vários algoritmos, sendo que, o que obteve melhores resultados foi a Random Forest [Breiman, 2001].

Dessa forma, o processo percorreu a seguinte sequência: pré-processamento dos textos, vetorização utilizando TF-IDF [Hiemstra, 2000] e geração do modelo utilizando Random Forest [Breiman, 2001]. Após os processos de otimização de hiper parâmetros obteve-se os seguintes resultados para as métricas avaliadas:

- ROC-AUC: 0,81
- AUPRC: 0,64

4.5. Modelo Estruturado

O processo de extração de variáveis gerou 127 variáveis. Alguns exemplos dessas variáveis extraídas dos textos são: quantidade de empresa com CNPJ no cadastro de empresas inadimplentes, quantidade de servidores públicos citados, quantidade de contratos vigentes, soma dos valores contratados, quantidade de pessoas beneficiadas por programas sociais citadas e etc.

Nem todas as variáveis eram relevantes para o processo de classificação das denúncias. Sendo assim, a primeira atividade realizada foi um estudo de influência das variáveis no processo de classificação.

Após a identificação da importância de cada uma das variáveis, ordenou-se essas variáveis em ordem decrescente de importância. Depois disso, gerou-se 127 modelos com diferentes números de variáveis, sendo que o modelo com menos variáveis possuía apenas uma (apenas a variável considerada mais relevante) e o com mais variáveis possuía 127 (número total de variáveis). Cada modelo sempre era gerado pelas variáveis mais relevantes.

Essa estratégia foi utilizada para identificar quantas variáveis seriam necessárias para a geração do modelo. O Figura 2 apresenta os gráficos obtidos pelos modelos gerados para as métricas avaliadas.

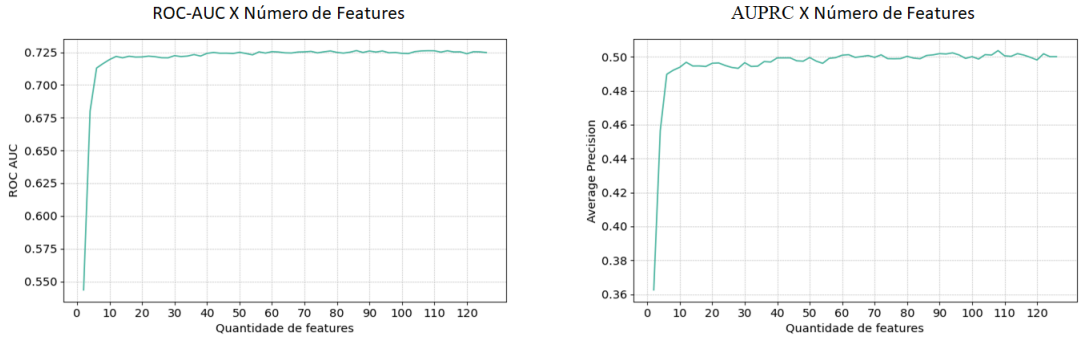

Figura 2. Desempenho dos Modelos por Número de Variáveis

Nos gráficos, o eixo horizontal representa o número de variáveis utilizadas pelos modelos. Sendo assim, os valores variam de 1 (número mínimo de variáveis utilizada por um modelo) a 127 (número máximo de variáveis utilizadas por um modelo – total de variáveis geradas). Já o eixo vertical representa o valor da métrica (ROC-AUC ou AUPRC) alcançada pelo modelo em questão.

Como pode ser observado, ambos os gráficos apresentam formas bem semelhantes. Os desempenhos dos modelos com menos variáveis são piores. Esse desempenho vai melhorando com a inserção de novas variáveis. No entanto, entre 10 e 20 variáveis o desempenho se estabiliza. Ou seja, a partir desse momento a adição de novas variáveis não interfere no desempenho do modelo.

Logo, conclui-se que 20 variáveis são suficientes para o processo de classificação. Dessa forma, a fim de deixar o modelo mais simples, selecionou-se apenas as 20 variáveis mais relevantes para o modelo final.

O modelo estruturado também foi gerado pelo algoritmo Random Forest [Breiman, 2001], e após a otimização dos parâmetros, obteve-se as seguintes métricas de avaliação.

- ROC-AUC: 0,79
- AUPRC: 0,56

4.6. Combinação dos Modelos

Uma vez que se tem os dois modelos gerados (um baseado nos dados estruturados e outro baseado nos dados textuais), o próximo passo é a combinação desses dois modelos. Para essa combinação optou-se pela utilização da média ponderada das previsões dos modelos individuais. Dessa forma, testou-se diferentes combinações entre os modelos a fim de se identificar aquela que apresenta melhores resultados. A Tabela 1 resume os valores obtidos.

Conforme pode ser verificado na Tabela 1, o modelo combinado 1 (que considera pesos iguais para os modelos originais) foi o que apresentou melhor resultado para a métrica principal (ROC-AUC) e segundo melhor resultado para a métrica secundária (AUPRC). Sendo assim, esse foi o modelo escolhido.

Tabela 1. Resumo dos valores obtidos

Identificação do Modelo	Peso dos modelos		AUPRC	ROC-AUC
	Estruturado	Textual		
Estruturado (original)	1	0	0,56	0,79
Textual (original)	0	1	0,64	0,81
Combinado 1	**1**	**1**	**0,65**	**0,84**
Combinado 2	2	1	0,64	0,83
Combinado 3	1	2	0,66	0,83

4.7. Análise

O último passo do estudo de caso é a comparação entre a performance do modelo textual e do modelo combinado, a fim de verificar se a aplicação da metodologia propiciou ganho de performance na classificação das denúncias.

Pela análise da Tabela 1 intui-se que o desempenho do modelo combinado superou o desempenho do modelo textual. Enquanto o modelo textual obteve o valor de 0,81 para a métrica ROC-AUC, o modelo combinado obteve 0,84. Da mesma forma, o modelo textual teve o valor de 0,64 para a métrica AUPRC contra 0,65 do modelo combinado. A fim de comprovar essa intuição, repetiu-se o experimento 1000 vezes e analisou-se a ROC_AUC para os modelos considerados, com o objetivo de validar se essa diferença se mantinha para outras divisões do conjunto de dados. Utilizou-se o teste Wilcoxon [1945] que comprovou, com nível de confiança de 95%, que o modelo combinado apresenta média superior à do modelo textual.

Essa mesma conclusão pode ser obtida a partir do gráfico apresentado na Figura 3. Essa Figura apresenta o gráfico da curva ROC-AUC para os dois modelos. O gráfico demonstra que a linha do modelo combinado fica acima da linha do modelo textual, o que evidencia que o modelo combinado apresenta desempenho melhor.

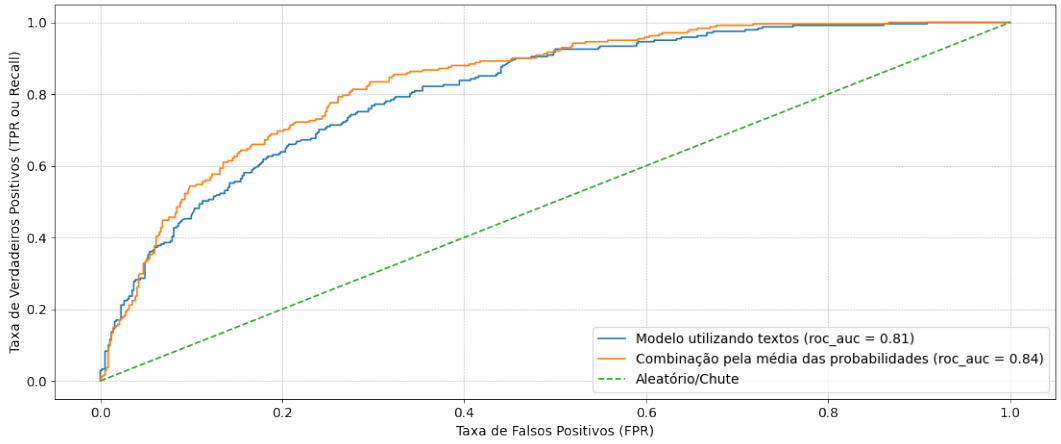

Figura 3. Gráfico da Área sob a curva ROC para os modelos Textual e Combinado

5. Conclusão

Este trabalho apresentou uma metodologia para extração e enriquecimento de variáveis em textos de denúncias. A metodologia propõe a identificação e extração de certos ele-

mentos dos textos das denúncias. Posteriormente é feita a validação desses elementos e busca-se outros elementos, derivados dos primeiros, utilizando para isso bases de dados externas. Por fim, a metodologia fornece um conjunto de dados estruturados que representa as denúncias.

Para validar a metodologia proposta foi executado um estudo de caso. Nesse estudo de caso foram gerados três modelos de classificação de denúncias: um obtido pelo processamento dos textos originais, outro obtido pelos dados estruturados (gerados pela metodologia proposta) e um terceiro que foi obtido pela combinação dos dois modelos anteriores.

Após as análises, constatou-se que o modelo combinado alcançou um desempenho superior ao modelo textual. Sendo assim, pode-se dizer que a hipótese: "se forem empregadas técnicas de extração e enriquecimento de variáveis em textos de denúncias, então, os resultados da classificação textual dessas denúncias podem melhorar" foi comprovada.

O conjunto de dados estruturados também é útil para outras atividades do tratamento das denúncias. Além disso, a identificação das variáveis estruturadas mais relevantes para o processo de classificação ajudou a área de negócio a entender quais seriam os fatores mais relevantes para determinar a aptidão de uma determinada denúncia.

Como trabalhos futuros, pretende-se melhorar as formas de representação das variáveis estruturadas, aprimorar o modelo textual de classificação de aptidão de denúncias e desenvolver modelos para classificar denúncia por área de apuração, sumarizar textos de denúncias, agrupar denúncias semelhantes e ranquear denúncias por dano potencial.

Referências

Alexandrino, M. and Paulo, V. (2006). *Direito administrativo*. Impetus.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

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.

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.

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

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

Hiemstra, D. (2000). A probabilistic justification for using tf x idf term weighting in information retrieval. *Int. J. Digit. Libr.*, 3(2):131–139.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

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.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, Z. (2019). A classification retrieval approach for english legal texts. In *2019 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, pages 220–223.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.

Liu, C.-z., Sheng, Y.-x., Wei, Z.-q., and Yang, Y.-Q. (2018). Research of text classification based on improved tf-idf algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, pages 218–222.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 2227–2237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radfort, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). (OpenAI Transformer): Improving Language Understanding by Generative Pre-Training. *OpenAI*, pages 1–10.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

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)*.

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.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

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.