Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Mariana Yukari Noguti

Aplicação de Técnicas de Classificação Textual na Predição de Áreas de Atuação do Ministério Público

Curitiba
2019

Mariana Yukari Noguti

Aplicação de Técnicas de Classificação Textual na Predição de Áreas de Atuação do Ministério Público

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Luiz Eduardo Soares Oliveira

Curitiba
2019

Aplicação de Técnicas de Classificação Textual na Predição de Áreas de Atuação do Ministério Público

Application of Text Classification Techniques in the Prediction of Public Prosecution Areas of Activity

Mariana Yukari Noguti[1]

[1]Ministério Público do Estado do Paraná, Sede Affonso Alves de Camargo, Bloco I, Curitiba, PR, Brasil *

Resumo

Observa-se nos últimos anos um crescimento no volume de pesquisas relativas a Processamento de Linguagem Natural (PLN). A utilização de redes neurais convolucionais e recorrentes em conjunto com técnicas de vetorização de palavras vem apresentando resultados promissores quando aplicadas a problemas de classificação textual, como análise de sentimentos e segmentação de documentos em tópicos. Neste artigo propõe-se o uso de técnicas de PLN na categorização de textos curtos, com o objetivo de classificar as descrições dos atendimentos realizados pelo Ministério Público do Paraná à população em uma das áreas de atuação da instituição. Buscou-se elaborar um modelo capaz de automatizar a rotulação dos atendimentos, reduzindo o tempo gasto com a seleção do atributo e a validação do cadastro, possibilitando a alocação de funcionários em demandas mais complexas. Foram utilizados métodos de extração de características textuais a partir de matrizes termo-documento e representações vetoriais. Na etapa classificatória foram apresentadas as performances obtidas por diferentes classificadores, dentre eles modelos lineares e *ensembles*, bem como algumas arquiteturas de redes neurais. Ao final, observou-se que o melhor resultado foi obtido através da representação vetorial de palavras com *Wang2Vec* associada à rede neural recorrente GRU, atingindo uma acurácia de 93% e *F1-Score* de 87,4% na classificação de doze categorias.

Palavras-chave: Processamento de Linguagem Natural, Classificação Textual, *Word Embeddings*, Redes Neurais Recorrentes

Abstract

In recent years, there has been an increase in the volume of research related to Natural Language Processing (NLP). The use of convolutional and recurrent neural networks together with word embedding techniques has presented promising results when applied to textual classification problems, such as sentiment analysis and topic segmentation of documents. This paper proposes the use of NLP techniques for categorization of short texts, with the purpose of classifying the descriptions of the services performed by the Public Prosecutor of Paraná to the population in one of the institution's areas of activity. It was intended to elaborate a model capable of automating the labeling of the attendances, reducing the time spent selecting the attribute and validating the register, allowing the allocation of employees in more complex demands. Methods of feature extraction from texts were compared by using document-term matrices and vector representations. In the classificatory stage were presented the performances obtained by different classifiers, among them linear models and ensembles, as well as some neural networks architectures. At the end, it was observed that the best result was obtained through vector representation of words with Wang2Vec associated with the GRU recurrent neural network, reaching an accuracy of 93% and F1-Score of 87.4% in the classification of twelve categories.

Keywords: Natural Language Processing, Text Classification, Word Embeddings, Recurrent Neural Networks

1. Introdução

O Ministério Público do Paraná (MPPR) é um órgão responsável pela representação dos interesses da sociedade, atuando diretamente em diversas áreas relacionadas aos direitos fundamentais da cidadania, tais como a defesa da saúde pública, do meio ambiente, do patrimônio público, dos direitos humanos, dentre outras. Uma de suas principais atribuições é a recepção de demandas através de atendimentos pessoais, telefonemas e denúncias, e posterior encaminhamento ao

*mynoguti@mppr.mp.br

setor mais adequado, sendo fundamental para isso a identificação da área de atuação responsável à resolução do caso. Desse modo, uma solicitação de medicamentos, por exemplo, deve ser remetida à Promotoria de Justiça na área da saúde, enquanto uma demanda por vaga em creche deve ser direcionada à área da educação. Atualmente, o MPPR conta com cerca de 470 unidades em todo o estado do Paraná que registram, em média, mais de dez mil atendimentos por mês. Todas as informações são cadastradas em um sistema eletrônico denominado PRO-MP e validadas por um Promotor de Justiça.

Uma amostra de 8.820 atendimentos revelou que aproximadamente 20% dos cadastros analisados apresentavam informações inconsistentes, em geral devido à associação incorreta da demanda, preenchida no sistema como um texto curto, em relação à área de atuação, um campo de múltipla escolha selecionado pelo usuário. A associação equivocada ocorre em razão de grande parte dos funcionários não possuírem formação na área do Direito, tornando a escolha da temática de atuação uma tarefa muitas vezes complexa. Além disso, a ausência de um protocolo unificado de preenchimento das informações no sistema reduz a uniformidade e a precisão dos dados cadastrados. Esse cenário reflete diretamente nas decisões tomadas pela administração, tendo em vista que muitas vezes se utiliza dessas informações para determinar a destinação de recursos humanos e materiais em áreas com maior volume de demanda. Por fim, a imprecisão no preenchimento da área de atuação responsável por um caso pode levar ao encaminhamento da solicitação à unidade incorreta, aumentando o tempo de tramitação até sua resolução.

Considerando o exposto acima, o presente artigo tem como objetivo confeccionar um modelo classificatório capaz de extrair informações textuais contidas nas descrições dos atendimentos, sugerindo automaticamente ao usuário a área de atuação correspondente. Pretende-se que a automatização dessa tarefa reduza a variabilidade interpretativa dos casos, bem como o tempo gasto pelo usuário na seleção do atributo no sistema, possibilitando a alocação desse tempo em demandas mais complexas. Além do ganho imediato no momento do cadastro, um classificador adequado possibilitaria a remoção da etapa de validação pelo Promotor de Justiça, encaminhando a demanda automaticamente ao setor adequado e diminuindo o tempo de tramitação dos casos. Por fim, o contínuo uso da ferramenta no sistema permitiria a longo prazo gerar estatísticas mais confiáveis das necessidades da população e, consequentemente, aprimorando a gestão administrativa.

A seguir serão reportados os resultados obtidos com a aplicação de técnicas de pré-processamento, extração de características e classificadores em uma base obtida do sistema PRO-MP relativo aos atendimentos realizados ao público. Serão apresentados os desafios em se trabalhar com a base, as avaliações do uso de uma matriz documento-termo (*Document-Term Matrix* ou DTM) comparativamente à representação vetorial de palavras e a avaliação de desempenho de classificadores clássicos, ensembles e redes neurais. Ao final será demonstrado que a representação obtida com *Wang2Vec* aplicada a redes neurais recorrentes apresentaram melhores resultados, especificamente com uso da denominada *Gated Recurrent Unit* (GRU), com obtenção de acurácia de 93% e *F1-Score* de 87,4% na tarefa de classificar doze diferentes áreas de atuação.

2. Trabalhos Relacionados

Estudos relacionados à classificação textual vêm aumentando nos últimos anos, motivados especialmente pela potencialidade na utilização de dados não estruturados contidos em documentos e registros textuais diversos. Particularmente na área do Direito, a aplicação de técnicas de *machine learning* e outras tecnologias é relativamente recente, tendo expandido em razão do surgimento de *lawtechs* ou *legaltechs*, empresas que propõem a conjunção da área do Direito e da tecnologia.

No que se refere a pesquisas específicas na área de Processamento de Linguagem Natural (PLN), é possível citar o artigo de Sulea et al. (2017) [1], o qual utilizou decisões judiciais da Suprema Corte da França e metadados diversos para predizer a área de Direito dos casos analisados, dentre outras tarefas. Para tanto, foram aplicados classificadores ensembles com extração de unigramas e bigramas para a representação textual. Mais similar ao caso estudado no presente artigo, o trabalho de Undavia et al. (2018) [2] propõe a classificação de documentos da Suprema Corte dos Estados Unidos entre quinze diferentes categorias, tendo comparado várias conjunções entre extração de características e classificadores, com obtenção de melhores resultados através da aplicação de redes neurais convolucionais (CNN) com representação vetorial utilizando a técnica *Word2Vec*. Um modelo similar foi aplicado em documentos recebidos pelo Supremo Tribunal Federal no

Brasil e relatado em Silva et al. (2018) [3], com obtenção de um classificador com uso de *embedding layer* e CNN aplicado a um problema de seis classes.

Outros métodos explorados neste trabalho são as aplicações de redes neurais recorrentes (RNN), em específico as denominadas *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU). Em relação ao tópico, pode-se encontrar em Wang et al. (2018) [4] um estudo comparativo entre LSTM e outros classificadores aplicados a análise de sentimentos em textos curtos. Foram utilizadas três diferentes bases, com melhor performance com uso de LSTM em casos em que havia grande volume de dados na partição de treino. Já no estudo de Yin e Kann (2017) [5], foi realizado um comparativo entre estruturas CNN e RNN em sete diferentes tarefas. Verificou-se no estudo que as RNNs apresentavam bons resultados em uma variedade de tarefas relacionadas a PLN. Ainda sobre o assunto, Wang et al. (2016) [6] realizou um comparativo entre arquiteturas de redes neurais recorrentes com uso de LSTM e GRU associadas a CNNs na classificação textual de três bases, tendo apresentado melhores resultados na conjunção dessas duas estruturas do que os demais modelos aplicados à mesma tarefa.

3. Materiais e Métodos

3.1. Materiais

A base de dados utilizada na tarefa de classificação foi obtida a partir dos registros de atendimentos no sistema PRO-MP, contendo 252.061 observações entre os anos de 2016 a 2019. Após uma análise preliminar, verificou-se a existência de elementos não classificados ou repetidos, bem como a existência de classificações incorretas e registros inadequados. Com o intuito de minimizar os reflexos desses problemas no treinamento do modelo, foi adicionada uma etapa intermediária de validação dos dados com auxílio de uma equipe do MPPR especializada na área de Direito. O objetivo era revisar a base e filtrar as observações inconsistentes, possibilitando reduzir a incidência de incorreções, de modo a assegurar maior eficácia do modelo e segurança nas predições obtidas. Em razão do pouco tempo disponível, foi possível analisar somente 19.702 registros, tendo sido validados deste montante o total de 15.263 observações, as quais foram consideradas no trabalho como base final para a construção do classificador.

Tendo em vista que algumas categorias apresentavam baixa representatividade em relação ao total de observações, optou-se por manter apenas classes com mais de duzentas observações na base de dados, resultando em um problema com doze categorias de classificação. Além disso, adotando o método comum de validação de modelos de Machine Learning, a base final foi dividida em três partições, com a proporção de dados igual a 80%, 10% e 10% do total de observações para as bases de treino, validação e teste, respectivamente. A distribuição de frequência das doze áreas de atuação nas três partições da base pode ser visualizada através do gráfico abaixo. Observa-se alto grau de desbalanceamento de classes, em especial para as áreas de Saúde e Família, com maior quantidade de atendimentos validados.

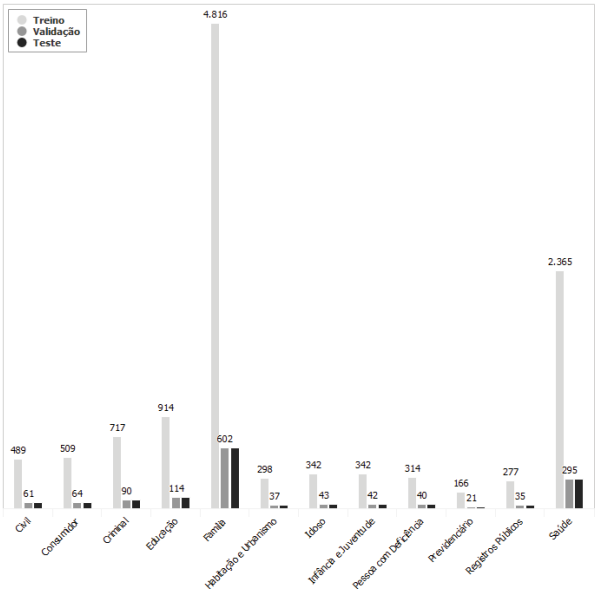


Figura 1: Distribuição das classes nas bases de dados

Em relação às técnicas de vetorização de palavras, foram obtidas alguns modelos pré-treinados em português e disponibilizados para uso através do artigo de Hartmann et al. (2017) [7]. Para o estudo foram selecionadas apenas as bases com tamanho vetorial de 50, 100 e 300 unidades, com uso de *Continuous Bag of Words* (CBoW) e *Skip-Gram* aplicados aos modelos *Word2Vec*, *Wang2Vec*, *FastText* e *Glove*. As bases acima foram treinadas com utilização de corpus genérico, não especializado na área do Direito.

Por fim, para treinamento de um modelo *Word2Vec* com vocabulário similar ao encontrado na base de classificação, foram utilizados dados da base de atendimento não validados e descrições de procedimentos internos do Ministério Público, igualmente registrados no PRO-MP, resultando em uma base contendo

922.588 sentenças para a criação de *word embeddings*, possibilitando a comparação do uso de bases genéricas e especializadas na tarefa de classificação textual.

3.2. Métodos

3.2.1. Pré-Processamento

A seguir estão listadas as técnicas utilizadas para pré-processamento textual neste artigo. As padronizações de palavras foram majoritariamente aplicadas com uso de expressões regulares, enquanto o processo de *lemmatization* e o modelo de *Part-of-Speech* (POS) *tagging* foram utilizados conforme constantes no pacote *spacy* do *python* para o idioma português.

1. Tokenização: divisão de sentenças em palavras;
2. *Lowercase*: padronização para letras minúsculas;
3. Remoção de pontuação;
4. POS *tagging*: classificação gramatical;
5. Padronização de numerais: transformação de números no algarismo zero;
6. Padronização de nomes próprios: transformação de nomes próprios no *token* nome_proprio a partir de uma lista de nomes em português coletada do IBGE [9] (dados tratados por Álvaro Justen/Brasil.IO);
7. *Lemmatization*: redução da inflexão das palavras em um núcleo comum;
8. Remoção de caracteres não-ASCII: exclusão de caracteres que não pertencem ao alfabeto ASCII.

Destaca-se que as etapas de pré-processamento foram aplicadas diferentemente de acordo com a seleção da técnica de extração de características dos textos, sendo descritas em cada caso na seção seguinte.

3.2.2. Extração de Características

Em relação à representação dos textos em termos numéricos, pode-se considerar a aplicação de duas técnicas distintas no presente estudo: a primeira com a criação de uma DTM e a segunda com aplicação de modelos de *word embeddings*.

Como a confecção de uma matriz termo-documento necessita de maior padronização das palavras para agrupar *tokens* correspondentes, foram utilizadas todas as etapas de pré-processamento descritas anteriormente, com o objetivo de reduzir a variabilidade dos termos encontrados nas descrições. Na etapa de aplicação do POS *tagging*, foram identificadas e mantidas apenas palavras cuja classe gramatical estivesse entre as seguintes: substantivo, verbo, advérbio e adjetivo, com o intuito de manter apenas as palavras consideradas mais relevantes na detecção de um assunto no atendimento. Da limpeza resultante, foi construída uma DTM com unigramas e bigramas contabilizados de acordo com a técnica *Term Frequency-Inverse Document Frequency* (TF-IDF), com linhas de corte configuradas para manter apenas *tokens* com mais de dez contagens e com menos de 90% de frequência considerando o total de documentos no *corpus*. A matriz resultante continha 6.428 termos, considerando todos os 11.549 documentos da base de treino.

No que se refere ao *word embedding*, foram aplicadas apenas algumas etapas do pré-processamento, tendo em vista que a técnica é capaz de extrair relações semânticas entre palavras, projetando-as em um espaço vetorial significativo e necessitando de menor quantidade de limpeza para uma boa representação do vocabulário associado ao *corpus* [8]. Como mencionado anteriormente, foram utilizadas duas representações vetoriais distintas, uma através de modelos pré-treinados de Hartmann et al. (2017) [7] e outra com aplicação do *Word2Vec* na base especializada do MPPR. Para cada caso, aplicou-se um processo de limpeza diferenciado.

Para o uso dos modelos pré-treinados, foi aplicado um *script* fornecidos pelos próprios autores para pré-processamento dos textos, a fim de se obter um estrutura semelhante e garantir a correspondência entre os *tokens* da base de classificação e dos modelos *embeddings* pré-treinados. Em linhas gerais, o pré-processamento consistia na tokenização da base, *lowercase*, remoção de estruturas especiais e de sentenças com menos de seis *tokens* e padronização de dígitos, palavras com pouca ocorrência, e-mails e *URLs*.

No segundo caso, foram aplicadas as etapas de tokenização, *lowercase*, remoção de pontuação, padronização de numerais e nomes próprios e remoção de caracteres não-ASCII. Para o treinamento dos modelos foi utilizado o pacote *gensim* do *python*, mantendo-se o tamanho vetorial igual a 50, 100 e 300, a fim de possibilitar a comparação com os modelos anteriores. Foram variados os tamanhos de janela de palavras e contagem mínima de aparições de palavras, bem como foram testadas a duas configurações disponíveis no pacote, o *CBoW* e o *Skip-Gram*.

3.2.3. Classificadores

Após a extração de características pelas técnicas descritas anteriormente, foram aplicados diversos classificadores com a finalidade de se obter a predição da área

de atuação a partir das descrições dos atendimentos. Para tanto, foram utilizados três grupos de classificadores, pontuados a seguir.

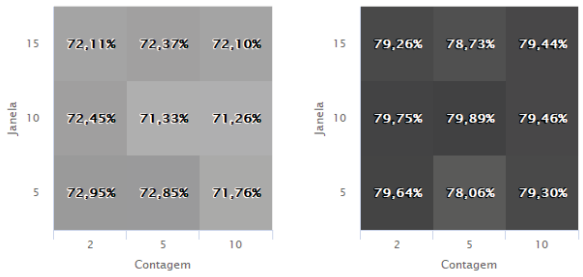
1. Classificadores Tradicionais:
   a) Regressão Logística
   b) *Support Vector Machine* (SVM)
2. Classificadores *Ensembles*:
   a) *Random Forest*
   b) *Gradient Boosting*
3. Redes Neurais:
   a) Redes Neurais Convolucionais (CNN)
   b) Redes Neurais Recorrentes (RNN)

Para a representação com TF-IDF foram aplicados os classificadores tradicionais e *ensembles*. No caso das representações com *embeddings* foram aplicadas as três categorias de classificadores, sendo que, para os dois primeiros grupos, a representação das descrições foi obtida através do vetor da média de todos os *tokens* que compunham a sentença, enquanto na aplicação de redes neurais foi utilizada a camada *embedding* do pacote *keras*, com padronização de tamanho de sentença igual a 200. Todas as arquiteturas foram configuradas no *keras* com *backend Tensorflow* e treinadas com número de épocas igual a 100 e auxílio de uma GPU.

Em razão do desbalanceamento de classes presente nos dados, foram utilizadas técnicas de balanceamento aleatório na base de treino, previamente à confecção dos classificadores, com uso do pacote *imblearn* do *python*. As duas técnicas aplicadas foram o *Random Under Sampler* (RUS), a qual obtém uma amostra aleatória balanceada das classes, com redução de todas as classes para a frequência da menor categoria observada, e *Random Over Sampler* (ROS), com obtenção de uma amostra aleatória com reposição balanceada das classes, a partir da reamostragem de elementos das classes menores até atingir a frequência da classe maior.

4. Resultados e Discussões

Com a finalidade de determinar a configuração dos parâmetros do *Word2Vec* na base especializada, optou-se por ajustar diversos modelos da referida técnica e testar um classificador simples e de rápido treinamento na base de validação, selecionando aquele que apresentasse melhor performance em termos de *F1-Score*. Desse modo, foram criados 54 modelos com uso das técnicas *CBow* e *Skip-Gram* e os valores de janela igual a 5, 10 e 15, contagem mínima das palavras igual 2, 5 e 10 e tamanho do vetor igual a 50, 100 e 300. Verificou-se inicialmente uma melhora nos resultados de modelos com vetor tamanho 300, tendo sido fixado esse parâmetro e gerado *heatmaps* dos *F1-Score* para visualização das demais configurações, conforme figuras abaixo.



(a) CBoW **(b)** Skip-Gram

Figura 2: *Heatmap* para *Word2Vec* com vetor tamanho 300

Em seguida realizou-se um teste similar nas bases pré-treinadas de textos genéricos. Neste caso foram testadas as variações dos modelos *Word2Vec*, *Wang2Vec*, *FastText* e *Glove*, variando-se o tamanho vetorial (50, 100 e 300) e as técnicas (*CBow* e *Skip-Gram*), exceto no caso do *Glove*, cuja metodologia não se utiliza das aludidas técnicas. Novamente foi verificado que o uso de vetor tamanho 300 apresentou melhora na performance do classificador. As representações dos valores *F1-Score* e acurácia em cada caso podem ser verificadas na tabela a seguir, considerando novamente a aplicação do classificador SVM e o vetor tamanho 300.

Tabela 1: Resultados da aplicação de SVM em configurações pré-treinadas de *embeddings*

Modelo	Técnica	Acurácia	F1-Score
Word2Vec	CBow	0,828	0,660
	Skip-Gram	0,846	0,698
Wang2Vec	CBow	0,864	**0,742**
	Skip-Gram	0,866	0,739
FastText	CBow	0,837	0,696
	Skip-Gram	0,864	0,739
Glove	-	0,850	0,722

Tendo em vista os resultados expostos acima e visando maximizar o *F1-Score*, foram selecionadas as configurações listadas em seguida para representação numérica dos textos na base de atendimentos:

1. TF-IDF: matriz com dimensão de 11.549 linhas (documentos) e 6.428 colunas (unigramas e bigramas);
2. *Word2Vec* com base especializada: aplicação da técnica *Skip-Gram* com vetor 300, contagem mínima de 5 e janela igual a 10;
3. *Word Embedding* com base genérica: modelo *Wang2Vec* com aplicação da técnica *CBoW* e vetor tamanho 300. Observa-se que o bom desempenho dessa técnica está de acordo com os resultados apontados em [7], no sentido de que o *Wang2Vec* apresenta boa performance em uma ampla variedade de tarefas de PLN.

Após a configuração das representações, foram aplicados os classificadores pontuados anteriormente. As tabelas abaixo apresentam os valores de acurácia e *F1-Score* na base de teste para cada modelo, divididas por grupos de classificadores. Em todos os casos foram testados o uso da base original, de RUS e ROS, a fim de se verificar a possibilidade de melhorias com aplicação de técnicas de balanceamento. Apenas o melhor resultado em termos de *F1-Score* será reportado a seguir, sendo descrito na coluna "Representação"caso tenha sido utilizada uma base com RUS ou ROS no lugar da original. A representação vetorial especializada será referida apenas como *Word2Vec*, enquanto a base genérica estará denominada como *Wang2Vec*.

Tabela 2: Resultados dos Classificadores Tradicionais na Base de Teste

Classificador	Representação	Acurácia	F1-Score
Regressão Logística	TF-IDF[b]	0,901	**0,835**
	Word2Vec[b]	0,865	0,789
	Wang2Vec[b]	0,805	0,714
SVM	TF-IDF[b]	0,899	0,829
	Word2Vec	0,890	0,791
	Wang2Vec[b]	0,821	0,730

[a] Utilização de RUS, [b] utilização de ROS

Tabela 3: Resultados dos Classificadores *Ensemble* na Base de Teste

Classificador	Representação	Acurácia	F1-Score
Random Forest	TF-IDF[b]	0,850	0,738
	Word2Vec[b]	0,825	0,656
	Wang2Vec[a]	0,618	0,523
Gradient Boosting	TF-IDF[b]	0,862	**0,782**
	Word2Vec[b]	0.859	0.751
	Wang2Vec[b]	0,791	0,655

[a] Utilização de RUS, [b] utilização de ROS

Tabela 4: Resultados do *Embedding Layer* e Redes Neurais na Base de Teste

Classificador	Representação	Acurácia	F1-Score
Conv 1D + MaxPooling 1D	Word2Vec	0,868	0,755
	Wang2Vec	0,863	0,746
LSTM	Word2Vec	0,916	0,838
	Wang2Vec	0,911	0,833
GRU	Word2Vec	0,922	0,854
	Wang2Vec	0,930	**0,874**
CNN + LSTM	Word2Vec	0,917	0,836
	Wang2Vec	0,911	0,825
CNN + GRU	Word2Vec	0,912	0,827
	Wang2Vec	0,884	0,774
Bi-LSTM	Word2Vec	0,909	0,821
	Wang2Vec	0,904	0,813
Bi-GRU	Word2Vec	0,918	0,841
	Wang2Vec	0,889	0,789

[a] Utilização de RUS, [b] utilização de ROS

A Tabela 2 demonstra que o uso de representação com DTM (TF-IDF) obteve performance superior em relação às demais no grupo dos classificadores tradicionais. O melhor resultado foi obtido com o classificador por Regressão Logística, sendo que o SVM também apresentou bons resultados. Em ambos os casos, foi verificado piora nos resultados com uso de representação com *word embeddings* na base genérica. Nota-se, ainda, que apenas a técnica SVM com *Word2Vec* apresentou melhores resultados com uso da base original, sendo que os demais foram positivamente afetados pela aplicação de ROS.

Em relação aos classificadores *ensemble* expostos na Tabela 3, não foi verificado aumento na performance em relação ao grupo anterior, de modo que, nas três representações dos textos testadas, foi notada uma diminuição do *F1-Score* em comparação com os modelos anteriores. Obteve-se uma melhora com aplicação de RUS com *Random Forest*, porém, ainda ficando significativamente abaixo do desempenho dos demais modelos analisadas até então. Nenhuma das combinações expostas nesta tabela apresentou bom desempenho com uso da base original.

Por fim, a Tabela 4 apresenta os resultados da aplicação do *embedding layer* do *keras* em conjunto com diferentes arquiteturas de redes neurais. A primeira estrutura foi baseada no trabalho de [3], com reconfigurações de alguns parâmetros que beneficiaram o desempenho do CNN na base utilizada. Ainda assim, os resultados foram medianos em relação aos demais e não superou o uso de classificadores tradicionais. Em seguida, foi utilizada a RNN LSTM, tendo apresentado bons resultados, com pequena melhora em relação ao desempenho da Regressão Logística. Um resultado

ainda melhor foi obtido através do uso da RNN GRU, sendo este classificador identificado como a melhor performance deste trabalho em ambas as representações textuais testadas. Ainda, foram analisadas outras configurações, com aplicação de LSTM e GRU bidirecionais, bem como de conjunção entre CNN e RNN, não tendo superado o uso isolado do GRU.

Tendo em vista que o melhor classificador foi obtido com a rede neural recorrente GRU associada à representação com *Wang2Vec*, apresenta-se a seguir o gráfico de treino e validação da acurácia e *F1-Score* nas 100 épocas treinadas, possibilitando a visualização do processo de construção do classificador.

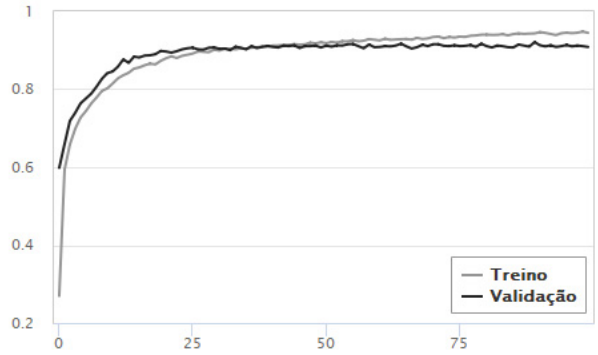


Figura 3: Acurácia durante o treinamento de 100 épocas nas bases de treino e validação

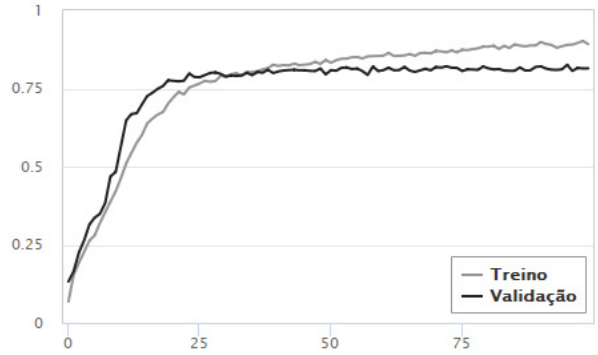


Figura 4: *F1-Score* durante o treinamento de 100 épocas nas bases de treino e validação

Pelas figuras acima, observa-se um crescimento considerável em ambas as métricas nas primeiras 25 épocas de treinamento, sendo que, após esse marco, a base de treino começa a apresentar sinais de *overfit*. Investigações mais aprofundadas nos parâmetros de *learning rate*, camada de *dropout* e diminuição de épocas foram realizadas, porém sem obter melhoria significativa no desempenho apresentado neste trabalho.

Mais abaixo são apresentadas as matrizes de confusão da base de testes gerada pelo RNN GRU, representando as doze áreas de atuação do Ministério Público classificadas neste estudo. A primeira configuração apresenta os valores brutos de classificações, enquanto a segunda demonstra a matriz com a proporção de acertos e erros em cada classe.

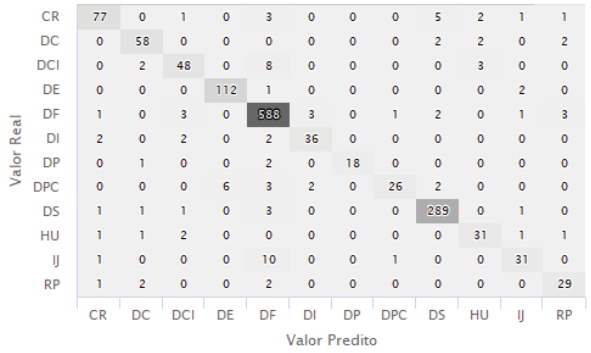


Figura 5: Matriz de confusão - número de ocorrências

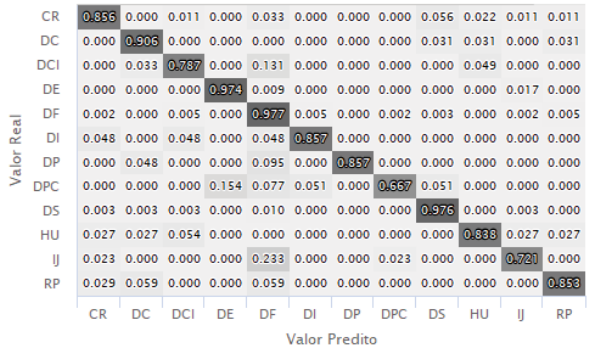


Figura 6: Matriz de confusão - proporção de ocorrências

Com o detalhamento das classes é possível verificar que os maiores índices de erros do modelo estão associados à área de atuação "Pessoa com Deficiência"e "Infância e Juventude", enquanto os maiores acertos se concentram nas área "Saúde"e "Família". Outro aspecto identificado foi o alto grau de confusão entre as classes "Cível"e "Infância e Juventude"com a classe "Família", podendo, em parte, ser explicado pelo fato dos vocabulários dessas áreas muitas vezes estarem correlacionados, sendo complexo até mesmo para um ser humano identificar a correta associação entre a demanda e o rótulo correspondente. Em todo caso, pode-se dizer que a utilização de redes neurais recorrentes apresenta resultados promissores, podendo ser expandida em outros problemas existentes no Ministério Público do Paraná.

5. Conclusão

Verificou-se no presente trabalho que a construção de classificadores textuais aplicados à área do Direito é possível, embora trabalhosa. O pré-processamento dos dados é uma etapa preliminar imprescindível à boa resolução de problemas, estando associada à qualidade das informações e, consequentemente, à precisão e confiabilidade do classificador.

Especialmente em se tratando de análise de textos, foi possível verificar que diferentes formas de representação podem apresentar bons resultados, devendo-se escolher pelo melhor método considerando o tempo para pré-processamento e os modelos a serem aplicados. Foi possível identificar que, embora as técnicas de POS *tagging* e *lemmatization* não estejam aperfeiçoadas ao idioma português, ainda assim podem ser utilizadas como ferramentas intermediárias e gerar bons resultados. Também foi notado que uma limpeza mais superficial com aplicação de técnicas de captura semântica apresentam igual ou melhores resultados, podendo ser mais trabalhosas quanto às etapas de treinamento e seleção paramétrica. Ainda sobre o assunto, foi possível identificar que o uso de vocabulários especializados apresenta ganhos em relação a uma base genérica apenas aplicados em conjunto com alguns modelos.

Quanto à estrutura dos dados, verificou-se que a aplicação de técnicas de balanceamento randômico obteve bons resultados em termos de melhoria do *F1-Score* nos classificadores mais simples testados. Para as redes neurais, em nenhuma das arquiteturas foi possível identificar ganhos na performance dos classificadores com uso de técnicas de balanceamento de classes. Pontua-se que esse tipo de modelo é altamente influenciado pelos hiperparâmetros selecionados [5], sendo necessária uma investigação mais aprofundada das configurações mais adequadas a cada base de dados e em cada uma das arquiteturas testadas.

Na etapa de classificação, foram obtidos bons resultados tanto em representações simples em conjunto com classificadores simples, como na representação e classificação realizadas através de redes neurais. No caso dos modelos *ensemble* foi notado um desempenho inferior em relação aos demais, podendo-se testar futuramente outras estruturas que se adequem melhor ao caso estudado. As redes neurais recorrentes obtiveram os melhores resultados nos experimentos, especialmente a GRU. Arquiteturas mais complexas não apresentaram ganho nas métricas avaliadas.

Pretende-se nas etapas seguintes aprofundar o presente estudo, no sentido de viabilizar a integração com um modelo capaz de identificar as palavras mais relevantes à predição de uma classe, inspirado no trabalho de Arras et al. (2017) [10]. Desse modo, será possível identificar, a partir da descrição de um caso, não apenas a área de atuação como também palavras-chave associadas. Por fim, os modelos serão inseridos no sistema eletrônico de registros do Ministério Público, visando à melhoria do fluxo de trabalho e das estatísticas geradas pela instituição, revertendo em benefícios concretos à população como um todo.

Referências

[1] O. Sulea, M. Zampieri, S. Malmasi, M. Vela, L.P. Dinu,J. van Genabith, *Exploring the use of text classification in the legal domain.* arXiv preprint arXiv:1710.09306 (2017).

[2] S. Undavia, A. Meyers, J.E. Ortega, *A Comparative Study of Classifying Legal Documents with Neural Networks*, Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE (2018), pp. 515-522.

[3] N. Correia Da Silva et al., *Document type classification for Brazil's supreme court using a convolutional neural network*, The Tenth International Conference on Forensic Computer Science and Cyber Law-ICoFCS (2018), pp. 7-11.

[4] J. Wang, T. Liu, X. Luo, L. Wang, *An LSTM Approach to Short Text Sentiment Classification with Word Embeddings*, The 2018 Conference on Computational Linguistics and Speech Processing (ROCLING 2018), pp. 214-223.

[5] W. Yin, K. Kann, M. Yu, H. Schütze, *Comparative study of cnn and rnn for natural language processing*, arXiv preprint arXiv:1702.01923 (2017).

[6] X. Wang, W. Jiang, Z. Luo, *Combination of convolutional and recurrent neural network for sentiment analysis of short texts*, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (2016), pp. 2428-2437.

[7] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, S. Aluisio, *Portuguese word embeddings: evaluating on word analogies and natural language tasks*, arXiv preprint arXiv:1708.06025 (2017).

[8] Y. Chen, B. Perozzi, R. Al-Rfou, S. Skiena, *The expressive power of word embeddings*, arXiv preprint arXiv:1301.3226 (2013).

[9] IBGE - Instituto Brasileiro de Geografia e Estatística, *Censo demográfico*, (2010). [Online]. Disponível em , acesso em: junho/2019.

[10] L. Arras, F. Horn, G. Montavon, K.R. Müller, W. Samek, *"What is relevant in a text document?": An interpretable machine learning approach*, PloS one, 12(8), e0181142 (2017).