Estudo de Abordagens para Classificação de Textos sobre Dúvidas Tributárias Utilizando Mineração de Texto

Rodrigo Dantas,
Departamento de Ciênmcia da Computação
Universidade doEstado do Rio de Janeiro -UERJ
Rio de Janeiro-Brasil
rodrigodc07@gmail.com

Karla Figueiredo
Departamento de Ciênmcia da Computação
Universidade doEstado do Rio de Janeiro -UERJ
Rio de Janeiro-Brasil
karlafigueiredo@ime.uerj.br

Leonardo Andrade
Secretaria de Fazenda do Estado do Rio de Janeiro- SEFAZ-RJ
Rio de Janeiro-Brasil
landrade@fazenda.rj.gov.br

***Abstract*— A SEFAZ-RJ possui um canal de atendimento pelo Sistema "Fale Conosco" usado para esclarecer dúvidas de contribuintes enviadas por e-mail. Com o distanciamento social, ocasionado pela COVID-19, a automatização das respostas para um maior volume de mensagens recebidas, tornou-se fundamental. Dessa forma, esse trabalho apresenta investigação de técnicas de Mineração de Texto, visando, em um primeiro momento, a classificação das mensagens de contribuintes a partir de técnicas de *Machine Learning/Deep Learning* para a automatização do processo de respostas aos contribuintes. Os resultados obtidos contribuíram para uma proposta de reformulação do formulário de dúvidas dos contribuintes, além de indicarem técnicas mais promissoras para se iniciar o processo de automatização do "Fale Conosco" da SEFAZ-RJ.**

Keywords—Mineração de Texto; Machine Learning; Deep Learning; Direito Tributário.

I. INTRODUCTION

O Sistema Tributário Nacional permite que a União, Estados e Municípios organizem suas próprias leis tributárias. Muitas vezes os entes federados criam ordenamentos jurídicos conflitantes entre si, causando um impacto jurídico negativo para as empresas. De acordo com [1], essa forma de estruturar um sistema tributário impõe um custo elevado para as empresas se manterem em conformidade fiscal.

Dada essa constatação, a Secretaria de Estado de Fazenda e Planejamento do Rio de Janeiro (SEFAZ-RJ) dispõe de um serviço de atendimento ao contribuinte, em que se propõe sanar dúvidas relativas à legislação tributária estadual, por meio de um canal de mensagem eletrônica. Nesse cenário, um especialista na área jurídica elabora as respostas às dúvidas postadas. Contudo, tal procedimento apresenta limitações, especialmente: i) a dificuldade na interpretação da legislação, trazendo grande demanda do corpo jurídico e técnico da SEFAZ-RJ por parte dos contribuintes; ii) alto custo de tempo e pessoal para realizar o trabalho manualmente; (iii) trabalho repetitivo, ou seja, muitos contribuintes apresentam dúvidas sobre um mesmo contexto, o que, eventualmente, poderia gerar respostas sem conformidade; e iv) somado a tudo isso, com a COVID-19, a necessidade de distanciamento social aumentou enormemente o número de mensagens, gerando atrasos que se tornam prejuízos aos cofres públicos, principalmente se o estado está submetido a um plano de recuperação fiscal.

Além disso, normalmente, os servidores da SEFAZ-RJ respondem manualmente aos questionamentos em até dois dias úteis. No entanto, durante a pandemia esses prazos tiveram que ser ampliados, devido ao maior número de mensagens recebidas.

Dessa forma, a área jurídica, em todos os seus ramos, em razão do volume de processos e, em geral, a ineficiência no andamento dos processos, vem atraindo muitos estudos de Processamento de Linguagem Natural a fim de simplificar alguns dos seus procedimentos, tal como a classificação de documentos jurídicos da suprema corte [2], [3] [4] e [5].

A área do direito tributário é uma das mais complexas por envolver dados textuais e numéricos (alíquotas, percentuais, valores, etc.). Sendo pouco explorada no processamento de linguagem natural, com raras exceções, como o trabalho de Ash and Marian (2019) [6] que exploram apenas a similaridade linguística, de forma empírica, entre pares de tratados em vigor em um determinado ano.

Dessa forma, dando início ao processo de automatização de respostas aos questionamentos dos contribuintes em português, foram investigadas soluções a partir de duas técnicas tradicionais da área de Mineração de Textos (MT), com *Support Vector Machine* (SVM) e redes *Long-Short Term Memory* (LSTM) explorando duas diferentes estratégias de estruturação da solução (direta e hierárquica) para, em um primeiro momento, realizar a classificação das dúvidas nas áreas tributárias estaduais. Assim este trabalho tem como objetivo específico investigar técnicas de MT, considerando textos em português no contexto específico do direito tributário, usando vetorizações diferentes e avaliação de bibliotecas de processamento de linguagem natural.

O restante do artigo está distribuído em mais quatro seções: na Seção II é apresentada uma breve introdução dos fundamentos técnicos necessários para melhor compreensão dos métodos e modelos desenvolvidos neste trabalho. A metodologia utilizada para solução do problema proposto, assim como sua aplicabilidade à esfera do problema é descrita na terceira Seção. Os estudos de casos são apresentados e os

resultados são discutidos na Seção IV, e, por fim, a última Seção apresenta as conclusões e perspectivas de novos trabalhos.

II. FUNDAMENTAÇÃO TEÓRICA

A. Support Vector Machine

O Support Vector Machine (SVM) [7] é um algoritmo transdutivo, baseado na Teoria da Aprendizagem Estatística, com resultados consagrados na área de Mineração de Texto para classificação de textos, sendo por esse motivo usado como *base line* para o trabalho.

B. Long Short Term Memory

Long-Short Term Memory (LSTM) [8] é um algoritmo de Redes Neurais Artificiais (RNA), que, para além das variáveis de entradas, também considera os estados anteriores da própria rede, de maneira que cada estado da rede influencie nos estados seguintes, sendo considerada como uma RNA recorrente. Além disso, as LSTMs são capazes de resolver o problema da dependência de longo prazo das RNAs recorrentes. Isso se deve a estrutura compostas por células LSTM, as quais possuem quatro unidades internas, que interagem com a informação que circula de forma diferenciada. Dada a característica da distribuição temporal associada às palavras que compõe uma frase ou texto, as Redes Neurais Recorrentes (RNR) têm sido largamente utilizadas em MT [9].

Assim, as LSTMs têm um estado adicional chamado estado da célula, com o qual o modelo pode remover ou adicionar informações cuidadosamente reguladas por estruturas denominadas gates (input, forget e outpt). Os gates são uma forma de, opcionalmente, deixar reter ou não as informações que passam pela rede neural.

O primeiro gate é o chamado de portão de esquecimento, no qual a célula decide quais informações serão descartadas do estado da célula. O segundo gate ou portão de entrada define quais informações serão adicionadas ao estado da célula. Este processo acontece em duas etapas: a primeira define quais valores serão atualizados, e a segunda cria uma lista de candidatos que podem ser incluídos. Em seguida, a saída dos dois estágios é combinada e é feita uma atualização do estado da célula. O terceiro gate é o portão de saída, no qual se determina quais informações serão apresentadas naquele momento pela rede LSTM [8].

C. Mineração de Textos/Processamento de Linguagem Natural

A Mineração de Texto (MT) é uma subárea da Descoberta de Conhecimento em Texto (KDT, do inglês Knowledge Discovery in Text) que consiste em extrair conhecimento, através de meios computacionais, sobre dados textuais não estruturados ou semiestruturados [9]. A mineração de textos possui duas abordagens: a análise estatística, que baseia-se principalmente no cálculo da frequência dos termos dentro do texto, e a análise semântica, que baseia-se no uso de algoritmos que efetuam a interpretação sintática e semântica do texto, buscando imitar a interpretação humana. As duas abordagens podem ser combinadas ou utilizadas de forma isolada [10].

Ambos os processos podem conter diversas etapas, porém quatro são comuns em todos os processos: coleta de documentos, pré-processamento, extração de conhecimento e avaliação e interpretação dos resultados [11].

O segundo estágio mencionado acima é o pré-processamento, o qual, a partir do texto coletado, prepara os dados, transformando-os em termos com alto valor agregado para o processamento dos algoritmos de aprendizado de máquina. Na etapa de extração de conhecimento os algoritmos extraem as características e padrões da base pré-processada, associando- às suas respectivas classes. Após essa etapa podem ser identificados padrões textuais que se repetirão nessas classes de textos, e com eles pode-se avaliar textos desconhecidos e indicar a qual classe pertence [10].

D. Vetorização

Na vetorização, uma das principais etapas do pré-processamento, as palavras são transformadas em números, e só então são consumidas pelos modelos de aprendizado de máquina. Genericamente, pode-se mencionar duas técnicas mais populares de vetorização: TF-IDF e *word embedding* [10]:

1) TF-IDF: visa transformar os termos em vetores de pesos. A métrica TF-IDF é uma fusão das métricas TF, que indica a frequência do termo em um documento (eq. 1) e IDF (Inverso da Frequência no Documento) aponta o quão raro é um termo entre os documentos (eq. 2). A métrica TF-IDF (eq. 3) avalia tanto a frequência de um termo quanto a sua relação de ocorrência em toda a coleção de documentos. O resultado deste processamento é uma matriz que representa a frequência relativa de cada palavra nos documentos avaliados.

$$\text{TF (i)} = \text{ocorrências do termo i/total de termos no documento} \quad (1)$$

$$IDF(i) = \log e^{(^{total\ de\ documentos}/_{nº de\ documentos\ com\ o\ termo\ i})} \quad (2)$$

$$TF - IDF = TF * IDF \quad (3)$$

2) *word embedding*: busca explorar a similaridade de certas palavras quando estão em um mesmo contexto [12]. A ideia é que a partir de uma rede neural do tipo *Multi-Layer Perceptron* (MLP), por exemplo, e uma base de dados textuais denominada *corpus*, possa se extrair as correlações entre os termos de acordo com o contexto, e, nesse caso, a posição relativa da palavra no texto também é considerada.

Para constituir um *corpus* são necessários conjuntos de textos/documentos contextualizados com a área de interesse, pois eles serão usados para que um modelo aprenda a relação entre as palavras. Assim, para se obter um *word embedding* capaz de detectar bem as relações entre as palavras, idealmente deve-se utilizar um *corpus* e um método capazes de extrair e representar bem as relações entre as palavras que serão vetorizadas.

O modelo adotado nesse trabalho foi o Word2Vec Continuous Bag-of-Words (CBOW), em que o modelo recebe uma janela de palavras extraída do texto, sem a palavra central, e deve prever a palavra omitida [12]. No resultado obtido por

esse método, os vetores que representam as palavras com sentidos parecidos ou correlatos estão próximos vetorialmente.

E. Fale Conosco da SEFAZ-RJ

A fonte dos dados (sob sigilo fiscal e por isso sem possibilidade de compartilhamento) processados nesse trabalho tem origem no atual sistema Fale Conosco de atendimento ao contribuinte da SEFAZ-RJ, por meio do qual o contribuinte pode sanar suas dúvidas. Nesse sistema o contribuinte se identifica por meio de e-mail, nome e CPF, que no caso deste trabalho foram omitidos. Em seguida, também deve preencher campos sobre a classificação de seus questionamentos: no primeiro campo deve identificar o imposto ou taxa ao qual a dúvida pertence, e num segundo campo, apontar o assunto, dentre um conjunto definido de assuntos, relativo ao imposto ou taxa. Esta associação visa facilitar o trabalho do corpo de servidores da SEFAZ-RJ no momento da elucidação da mesma, e também para referências futuras a outros problemas fazendários. No entanto, muitas vezes os preenchimentos desses campos não correspondem ao conteúdo das mensagens, o que invalida a possibilidade de respostas automáticas apenas pela identificação do tributo e assunto.

III. METODOLOGIA

Visando investigar as opções iniciais para a automatização das respostas às dúvidas dos contribuintes, duas estratégias foram usadas para construir modelos para inferência de assuntos relativos aos tributos. Assim, a partir da identificação do tipo de assunto, pode se planejar o desenvolvimento de modelos que, mais especificamente, podem aprender a responder adequadamente cada assunto. Na primeira abordagem proposta, denominada direta (Figura 1), um único modelo de classificador irá classificar a dúvida (texto da mensagem) em um assunto, considerando os assuntos relacionados a todos os impostos e taxas arrecadados pela SEFAZ-RJ. Em uma segunda abordagem, que consiste em construir um modelo hierárquico, o texto deve ser inicialmente classificado entre um dos impostos ou taxa, para em seguida ser novamente classificado por outro classificador especializado apenas nos assuntos daquele imposto ou taxa, conforme descrito na Figura 2.

Para os dois tipos de abordagem (direta ou hierárquica), as classificações foram realizadas por dois algoritmos: um classificador utilizando SVM e um classificador utilizando LSTM, com o objetivo de comparar e analisar os desempenhos dos algoritmos e parâmetros. O modelo baseado em SVM foi escolhido, entre outros algoritmos de *Machine Learning*, por ser um algoritmo que apresenta bom desempenho para bases com muitos atributos e poucos registros e bases com algum grau de desbalanceamento entre as classes [13]. Para essas classificações as bases de dados foram estratificadas e separadas em duas partes na proporção: 90% da base de dados para treinamento dos modelos (validação cruzada = 10 k-folds) e os 10% restantes para testar os modelos.

Para a abordagem hierárquica, foi necessário realizar a inferência de tributos, e nesse caso o modelo deveria classificar as dúvidas em um dos três tipos de impostos e a taxas aplicadas pelo Estado do Rio de Janeiro: ICMS (Imposto sobre operações relativas à Circulação de Mercadorias e sobre prestações de Serviços de transporte interestadual e intermunicipal e de comunicação), IPVA (Imposto sobre Propriedades de Veículos Automotores), ITD (Imposto sobre Transmissão causa-mortis e Doações de quaisquer bens e direitos) ou Taxas Estaduais.

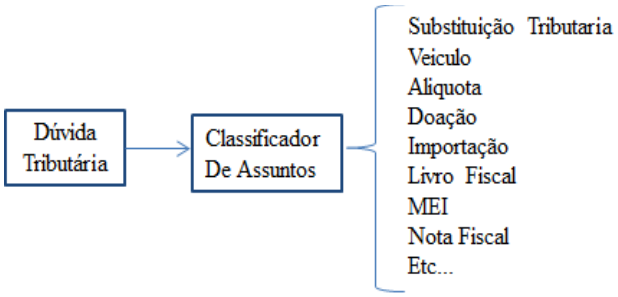

Fig. 1. Modelo Classificação Direta

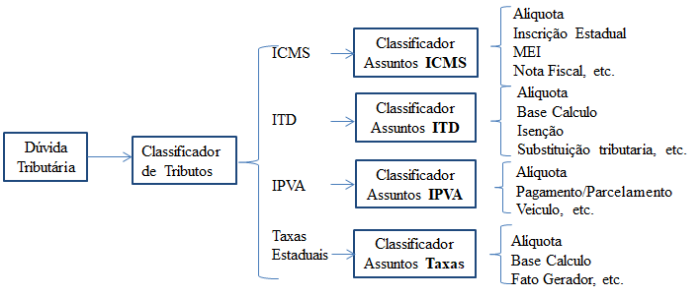

Fig. 2. Modelo Classificação Hierárquico

IV. ESTUDO DE CASO

A. Base de Dados da SEFAZ-RJ

O trabalho foi, nesse primeiro momento, focado nas dúvidas dos contribuintes coletadas pelo sistema de auxílio ao contribuinte da SEFAZ-RJ, que devido às mudanças recorrentes na legislação tributária, ficou restrita ao ano de 2018. Outro motivo para a restrição do volume da base foi a necessidade de validação manual dos rótulos (impostos e taxas e seus assuntos), registrados pelos contribuintes no momento da postagem da dúvida, pelos auditores e fazendários.

A base de dados é composta pelos atributos: data da solicitação, assunto, tributo, atendente, revisor, protocolo, dúvida e resposta. Originalmente os registros estão distribuídos em três impostos e taxas, e em seus respectivos 56 assuntos.

O total de documentos (dúvidas) por impostos/taxas é: IPVA=281, ICMS= 10595, ITD= 170 e Taxas=36. Devido a quantidades de dúvidas diferentes para cada assunto, os assuntos que tinham menos de 100 dúvidas foram removidos, porque com baixo número de dúvidas, o conhecimento extraído não se mostrou, em testes preliminares, suficiente para generalizar e mapear as características da dúvida associadas a estes assuntos. Assim, os assuntos foram reduzidos a 25 assuntos, enumerados abaixo e indicados na Figura 3, restando apenas os assuntos do ICMS.

- Assuntos do ICMS: Alíquota, Base Cálculo, Benefício Fiscal, CFOP, Conhecimento Transporte, Crédito, CTE, Diferencial Alíquota Rifa ECF, EFD, Fato Gerador, FECP, Importação, Inscrição Estadual, Isenção, Livro Fiscal, NFC-e, NF-e, Nota Fiscal, Pagamento, Parcelamento, Restituição,

Saldo Credor, Simples Nacional, SPED e Substituição Tributária;

Além disso, por conta da baixa quantidade de dúvidas para inferência de impostos/taxas também foram removidas as dúvidas sobre taxas estaduais.

Dessa forma, na primeira etapa da abordagem hierárquica, a classificação dos impostos (ICMS, IPVA e ITD), foi mantida com a distribuição de documentos indicado acima, mesmo com o desbalanceamento entre as classes [13]. O objetivo foi avaliar a capacidade de aprendizado do contexto do problema com esse conjunto de dados, em um primeiro momento, sem explorar técnicas de balanceamento. Assim, os assuntos relativos aos impostos IPVA e ITD tinham, além do desbalanceamento, um número muito pequeno de dúvidas por assunto. De qualquer maneira, as dúvidas dos assuntos dos impostos ICMS, IPVA e ITD foram reunidas, respectivamente em cada um deles, para compor os documentos que seriam relacionados a esses impostos para o treinamento da classificação da primeira etapa da abordagem hierárquica. A classificação de assuntos, na segunda etapa da estratégia hierárquica, focou somente nos assuntos do ICMS, pois, conforme já descrito acima, foi o imposto que tinha um volume de dúvidas por assunto que permitiu o processo de aprendizado.

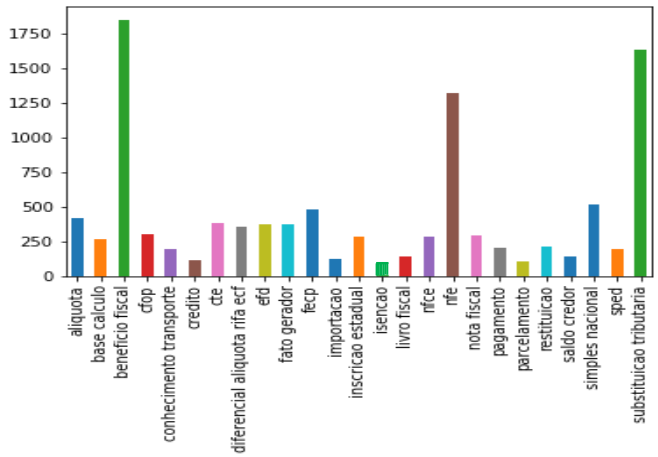

Fig. 3. Distribuição da quantidade de dúvidas por assuntos para ICMS

B. *Pré-Processamento dos documentos*

Os dados utilizados em ambos os classificadores (SVM e LSTM) foram transformados pelas seguintes etapas de pré-processamento: "tokenização", *case folding*, remoção de acentos, remoção de dígitos, remoção de pontuação, remoção de *stopwords*, correção ortográfica, *stemming* e vetorização [10]. Destaca-se que para o SVM foi utilizada a vetorização com o TF-IDF, configurando a abordagem estatística de Mineração de Texto e para LSTM foi usada a técnica de *word embedding*, configurando a abordagem semântica, embora ambas as vetorizações pudessem ser utilizadas nesses algoritmos. O objetivo dessa definição foi avaliar se com um algoritmo e vetorização mais simples, baseado em frequência de termos (TF-IDF), seria possível atingir os objetivos desejados. Além disso, a métrica TF-IDF contorna o problema do *corpus* específico que o *word embedding* acarreta.

Para a correção ortográfica foi utilizado o dicionário de Português (Brasil) da ferramenta WinEdt [14]. Além disso, foram adicionadas 70 palavras a este dicionário, como por exemplo: base calculo, beneficio fiscal, alíquota, fato gerador, livro fiscal, mei, regime especial, sepd, simples nacional e sintegra, a fim de complementar o dicionário com palavras presentes no vocabulário tributário fiscal estadual.

C. *Ajustes dos Modelos*

Inicialmente foram investigados os parâmetros mais adequados para os modelos SVM e LSTM com metodologia direta.

Para definir os parâmetros do modelo SVM, utilizando a abordagem *one-against-all*, construído na plataforma scikit-learn, onde foram testados diferentes valores de custo (C) e funções kernel e seus respectivos parâmetros.

A Tabela I aponta a acurácia e o F1-score dos melhores resultados, por kernel avaliado, obtidos utilizando os dados da base de validação para as propostas de abordagens: direta e hierárquica, apresentadas na Seção III, para a classificação de assuntos por SVM.

Na avaliação da Tabela I se observa muitos resultados repetidos, porém, como esperado, a abordagem hierárquica obteve melhores resultados para o algoritmo SVM. Destacando-se em negrito a opção que contem a função de kernel ***sigmoid*** e custo 1000, tanto para a abordagem direta como para a hierárquica.

TABELA I. F1-SCORE E ACURÁCIA DE VALIDAÇÃO PARA CLASSIFICAÇÃO DE ASSUNTOS USANDO SVM, COM ABORDAGEM DIRETA E HIERÁRQUICA, PARA DIFERENTES PARÂMETROS

	Custo (C)	**Função Kernel**	**Acurácia**	**F1**
Direta	10	rbf*	0,1656	0,0114
	10	*sigmoid*	0,1656	0,0114
	10	polinomial	0,1656	0,0114
	100	rbf*	0,3182	0,0631
	100	*sigmoid*	0,1674	0,0124
	100	polinomial	0,1656	0,0114
	1000	rbf*	0,3182	0,0631
	1000	***sigmoid***	**0,4960**	**0,3072**
	1000	polinomial	0,1656	0,0114
Hierárquica	10	rbf*	0,1819	0,0139
	10	*sigmoid*	0,1819	0,0139
	10	polinomial	0,3327	0,0653
	100	rbf*	0,3352	0,0653
	100	*sigmoid*	0,1833	0,0147
	100	polinomial	0,1819	0,0139
	1000	rbf*	0,3352	0,0653
	1000	***sigmoid***	**0,5751**	**0,3572**
	1000	polinomial	0,1819	0,0139

*rbf: radial base function

O classificador LSTM foi construído na plataforma Keras. Os modelos usaram um histórico de entrada de até 250 palavras (valor mínimo observado nas mensagens postadas pelos contribuintes), entropia cruzada binária como função de perda, algoritmo de otimização Adam, *earlystopping*, além de utilizar *mini-batches* de tamanho igual a 128, um máximo de 100 épocas, função de ativação Linear Retificada (ReLu) nas camadas intermediárias e *sigmoid* na saída e 0.5 de *dropout*. Além disso, foram avaliadas diferentes dimensões e funções de ativação nas camadas intermediárias.

A Tabela II apresenta os resultados médios das acurácias e F1-Score de validação cruzada, em um resumo das parametrizações exploradas.

Tanto para o algoritmo SVM (Tabela I) como LSTM (Tabela II), as classificações de assuntos sob a abordagem hierárquica se saíram melhor. Comparando as duas tabelas I e II também se pode considerar que o algoritmo LSTM foi superior ao SVM. No entanto, não apresentou diferença significativa na comparação entre a abordagem direta e hierárquica. A explicação pode estar no corte dos assuntos e das taxas estaduais devido ao baixo volume de documentos, reduzindo a vantagem da abordagem hierárquica. Assim, a base de teste foi avaliada na abordagem hierárquica com o algoritmo LSTM.

TABELA II. F1-SCORE E ACURÁCIA DE VALIDAÇÃO PARA CLASSIFICAÇÃO DE ASSUNTOS USANDO LSTM COM ABORDAGEM DIRETA E HIERÁRQUICA

	Função de Ativação	Quantidade de células	Acurácia	F1
Direta	relu	128	0. 6243	0,5826
	relu	256	0,4929	0, 4779
	relu	**512**	**0, 7206**	**0, 6929**
	sigmoid	128	0, 6745	0, 6216
	sigmoid	256	0, 6933	0, 6628
	sigmoid	512	0, 7146	0, 6891
	exponencial	128	0, 4588	0, 3432
	exponencial	256	0, 4637	0, 3491
	exponencial	512	0, 6817	0, 64778
Hierárquica	relu	128	0, 6870	0, 6432
	relu	256	0, 6991	0, 6603
	relu	**512**	**0, 7224**	**0, 6949**
	sigmoid	128	0, 6552	0, 6099
	sigmoid	256	0, 6677	0, 6230
	sigmoid	512	0, 6619	0, 6280
	exponencial	128	0, 4865	0, 3865
	exponencial	256	0, 4707	0, 3614
	exponencial	512	0, 7059	0, 6683

D. Testes dos Modelos

Após o treinamento dos algoritmos, considerando as investigações de seus parâmetros, na Tabela III são apresentadas as acurácias e F1-score por classe de saída para a etapa 1 da abordagem hierárquica de classificação de impostos do conjunto de teste. A quantidade de documentos usados no **treinamento das redes LSTM** também foi incluída na Tabela III com o objetivo de indicar o desbalanceamento e, consequente, desempenho dos resultados por imposto. Destaca-se que, apesar disso, o IPVA não apresentou um resultado tão ruim quanto o ITD, mesmo em desvantagem numérica de documento com relação ao ICMS. Assim, a acurácia total do algoritmo LSTM para esta etapa de classificação de impostos da abordagem hierárquica foi 94,23% e a Figura 4 apresenta a matriz de confusão percentual para classificação de impostos para o conjunto teste.

Na Tabela IV são apresentadas as acurácias e F1-scores dos melhores modelos LSTM (destacados na Tabela 2), utilizando a abordagem hierárquica, para classificação de assuntos. Lembrando que esta classificação envolve 25 diferentes assuntos.

TABELA III. RESULTADO DE CLASSIFICAÇÃO HIERÁRQUICA DE IMPOSTOS (ETAPA 1 DO MODELO HIERÁRQUICO) – BASE TESTE

Impostos	Quantidade de dúvidas	Acurácia	F1-score
ICMS	7952	0,8133	0,7813
IPVA	195	0,9100	0,8508
ITD	131	0,8367	0,6994

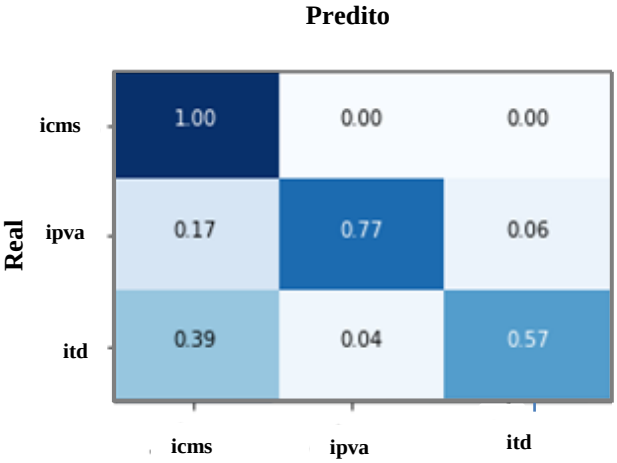

Fig. 4. Matriz de Confusão para o Classificação de Impostos (Etapa 1 Modelo Hierárquico) – Base Teste

TABELA IV. RESULTADO DE CLASSIFICAÇÃO HIERÁRQUICA DE ASSUNTOS PARA A BASE TESTE – LSTM

Assuntos (25 assuntos) Abordagem Hierárquica	
Acurácia	**F1 Score**
0,5247	0,4945

Devido ao reduzido espaço e grande quantidade de assuntos, é inviável apresentar a matriz de confusão completa com o s resultados da classificação dos 25 assuntos. Por esse motivo, foram selecionados os assuntos com maior incidência de dúvidas (mais de 300 dúvidas) para serem apresentados.

A partir da Figura 5 observa-se que o maior erro de classificação ocorreu para as dúvidas sobre alíquota, que foram erroneamente classificadas como dúvidas sobre benefício fiscal. Nesse caso pode se entender que quando um benefício fiscal é proposto, em geral, altera-se o padrão da alíquota e/ou a base de cálculo/fato gerador. Assim, faz sentido que as dúvidas sobre benefício fiscal incluam termos relativos à alíquota e/ou fato gerador.

A seguir, são apresentadas argumentações sobre alguns erros na classificação dos assuntos da base teste observadas na matriz de confusão (Figura 5):

- no momento de preenchimento de uma nota fiscal eletrônica, os produtos inseridos precisam ser identificados com os códigos fiscais de operações e prestações de serviços (CFOP), por isso é comum encontrar termos relacionados ao código fiscal de operações e prestações de serviço (CFOP) em dúvidas sobre nota fiscal eletrônica (NFE).
- quanto ao assunto "fato gerador", observa-se que ele é o que apresenta menor métrica de acerto. Isso pode ser explicado porque o **fato gerador** é hipótese de incidência tributária que **define um tributo**, ou seja, caso a hipótese descrita em lei seja verdadeira, determina a obrigação tributária. Assim, os termos presentes nesse assunto podem estar presentes em diversos assuntos.
- o **diferencial alíquota** é um tipo de **alíquota**, por isso as dúvidas que envolvem esse assunto são confundidas com o assunto "alíquota". Vendas interestaduais podem ser exemplos da aplicação desse tipo de alíquota.
- o assunto escrituração fiscal digital (EFD) também foi algumas vezes confundido com nota fiscal eletrônica (NFE). A legislação tributária exige que mensalmente as NFEs sejam relacionadas no em um arquivo eletrônico (EFD), com regras próprias de preenchimento. Assim, é razoável que as dúvidas dos contribuintes sobre esses assuntos versem sobre os mesmos termos.

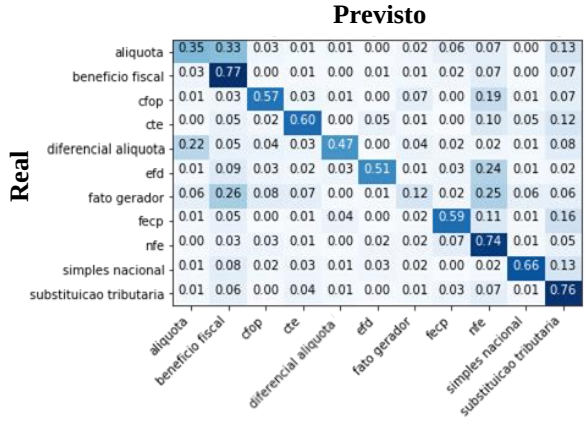

Fig. 5. Matriz de Confusão para o Classificação de Assuntos (Etapa 2 Modelo Hierárquico) considerado os assuntos com mais de 300 menções na Base Teste

Além dessas observações, acrescenta-se que quanto ao desempenho da classificação de assuntos, seja por meio do modelo direto ou hierárquico, durante o pré-processamento o corretor ortográfico alterou 11% das palavras presentes nas bases. Isso demostra a grande quantidade de erros que pode ocorrer quando usuários preenchem campos de texto livre. Registra-se também que o *word embedding* (CBOW) (com vetores de dimensão igual a 50) utilizado só foi aplicado em 53% das palavras presentes, retratando a peculiaridade da base de dados, onde as dúvidas têm um vocabulário muito específico, relativo ao direito tributário. Assim, pode-se perceber que o vocabulário das dúvidas não tinham perfeita correspondência com um *word embedding*, gerado a partir de *corpus* de contexto geral. Isso indica que se a vetorização tivesse como base um *word embedding* específico para o contexto tributário estadual, esses resultados poderiam ser melhores.

A partir dos tipos de assuntos descritos na Seção IV.A e IV.D, fica evidente a mistura de termos que existe entre os assuntos do ICMS. Sugerindo que os assuntos possam ser mais bem estabelecidos, de forma a se tentar gerar menos confusão por parte dos contribuintes, como também ajudando o processo de aprendizado dos modelos.

Também se constata que, a despeito do *corpus* específico do problema não estar totalmente coberto pelo processo de vetorização usado nas redes LSTM, os melhores resultados são obtidos pelo algoritmo LSTM, em detrimento do TF-IDF, que usa apenas os termos dos próprios documentos para gerar a vetorização, sem depender de corpus externo. Também é notável que, mesmo com o desbalanceamento das classes de impostos (ICMS, IPVA e ITD), os resultados de IPVA e ITD foram considerados bons.

V. Conclusão e Trabalhos Futuros

Este trabalho investigou duas abordagens de classificação para textos em português, contendo dúvidas tributárias do Estado do Rio de Janeiro (Fale Conosco da SEFAZ-RJ). As dúvidas são rotuladas pelos contribuintes em três tipos de impostos ou taxas e em seus respectivos assuntos, totalizando 56 diferentes tipos de assuntos. O trabalho investigou duas alternativas para classificação de assuntos, uma primeira abordagem direta e uma segunda hierárquica, usando dois tipos de algoritmos: SVM e LSTM.

Foi necessário reduzir o número de assuntos, devido à baixa quantidade de dúvidas em determinados assuntos. Este é um problema que está ajudando a mobilizar novos esforços para o aumento do conjunto de documentos a ser utilizado.

Verificou-se que o SVM possui razoável propriedade de generalização, além de se mostrar adequado para bases de dados com muitos atributos e poucos registros desbalanceados [13]. O LSTM, apesar da boa acurácia obtida, teve dificuldade durante a classificação de alguns assuntos. Isso pode ser atribuído a dois fatores: a baixa quantidade de dúvidas de alguns assuntos, dificultando a aprendizagem dos modelos, que por natureza (*Deep Learning*) exigem uma quantidade maior de informação, e a grande similaridade semântica entre alguns assuntos, o que leva a uma diminuição no índice de acertos nesses assuntos. Esta última questão deve provocar uma nova proposta de conjuntos de assuntos, que permita ao contribuinte a escolher opções mais assertivas com relação às suas dúvidas.

Assim, um primeiro campo a ser futuramente explorado é o aumento da base de dados de dúvidas, mas que precisa ser manualmente verificada pelos auditores. Além disso, deve-se levar em consideração o uso de um *corpus* específico do contexto para realizar o *word embedding* do modelo de LSTM,

pois conforme indicado na Seção IV.D, muitas palavras não foram vetorizadas, devido ao fato do *corpus* utilizado não totalmente aderente ao escopo do direito tributário. Por fim, pode-se mencionar também algoritmos para pergunta-resposta como opção de próximos passos, utilizando tecnologia baseada em *Transformers* (BERT [15] e suas evoluções [16]), além do *framework Text-To-Text Transfer Transformer* (T5)[17].

Conclui-se, assim, que mesmo considerando as dificuldades encontradas, os resultados obtidos se revelaram promissores para um futuro processo de automatização da elucidação de dúvidas dos contribuintes, tendo contribuído desde orientações para definição de conjuntos de assuntos a ser selecionado pelo contribuinte, que possuam menos misturas de termos, até indicações de abordagens de solução para a correta identificação da dúvida.

REFERÊNCIAS

[1] B. Appy, Por que o sistema tributário brasileiro precisa ser reformado. Interesse Nacional, 8(31), pp.65-81, 2015.

[2] P. Casanovas, M. Palmirani, S. Peroni, T. van Engers, and F. Vitali, Semantic web for the legal domain: the next step. Semantic web, 7(3), pp. 213-227, 2016.

[3] R. Dale, Law and Word Order: NLP in Legal Tech, Published online by Cambridge Univ.Press. 2018, DOI: https://doi.org/10.1017/S1351324918000475

[4] L. Robaldo, S. Villata, A. Wyner, et al. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts". Artificial Intelligence. Law 27, pp. 113–115, 2019.

[5] F. Fagan, Natural Language Processing for Lawyers and Judges, Michigan Law Review, Forthcoming. 2020 Available at SSRN: http://dx.doi.org/10.2139/ssrn.3564966

[6] E. Ash and O. Marian, The Making of International Tax Law: Empirical Evidence from Natural Language Processing. UC Irvine School of Law Research Paper No. 2019-02. Available at SSRN: https://ssrn.com/abstract=3314310

[7] C. Cortes and V. Vapnik, Support-vector networks. Machine learning, 20(3), pp. 273-297. 1995.

[8] S. Hochreiter and J. Schmidhuber, Long short-term memory. Neural computation, 9(8), pp. 1735-1780, 1997.

[9] D. Jurafsky and J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Prentice Hall PTR, 2000.

[10] A. Zhang, Z. Lipton, M. Li, and A. Smola, Dive into Deep Learning. 2020. https://d2l.ai.

[11] A. Pezzini, Mineração de Textos: Conceito, Processo e Aplicações, 2017. DOI: 10.5965/2316419005082016058

[12] T. Mikolov, W. Yih, and G. Zweig, Linguistic regularities in continuous space word representations. In Proc. Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Techno. pp. 746-751, 2013.

[13] descaracterizado, Titulo oculto. In: Congresso, XXX p. aaa-zzz, ano.

[14] A. Simonic, WinEdt Unicode Dictionaries. In: WinEdt Unicode Dictionaries. [S. l.], 23 fev. 2016. Disponível em: http://www.winedt.org/dict.html. Acesso em: 14 abr. 2019.

[15] J. Devlin, M-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 2018 arXiv:1810.04805v2 [cs.CL].

[16] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, "Efficient Transformers: A Survey", ArXiv abs/2009.06732, 2020.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, v. 21, 140, pp.1-67, http://jmlr.org/papers/v21/20-074.html