PROPOSTA TÉCNICA

Chamamento Público 001/2022 para Encomenda Tecnológica (ETEC)
do Tribunal de Contas da União (TCU)

25 de março de 2022
Brasil

Sumário

1. Introdução

Trata-se de proposta formulada pelo consórcio a ser formado pelas organizações RBCIP-FABWORK-N2VEC para Encomenda Tecnológica de um módulo de Instrução Assistida por Inteligência Artificial, a ser incorporado à solução de Instrução Assistida do TCU, conforme solicitado por meio do Edital de Chamamento Público para Encomenda Tecnológica de Instrução Assistida por Inteligência Artificial, disponível em link.

O consórcio a ser formado RBCIP-FAPWORK-N2VEC foi formalizado por meio do termo de compromisso de constituição de consórcio, tendo como empresa líder a Rede Brasileira de Certificação, Pesquisa e Inovação (RBCIP). A RBCIP é uma associação civil com personalidade jurídica de direito privado, sem fins econômicos, estatutariamente e legalmente (Lei nº 10.973/04 e 13.243/16) enquadrada como instituição científica, tecnológica e de inovação (ICT) e sediada no Distrito Federal. No âmbito do Distrito Federal, a RBCIP é credenciada na Fundação de Apoio à Pesquisa do Distrito Federal como Organização da Sociedade Civil e goza dos benefícios previstos na lei Lei nº 8.010, de 29 de março de 1990, alterada pela Lei nº 10.964, de 28 de outubro de 2004. Tem como missão fomentar e promover o ensino, a pesquisa científica, o desenvolvimento tecnológico e o desenvolvimento institucional. A RBCIP mantém a transparência de suas informações por meio de seu site www.rbcip.org.

A FABWORK é uma *startup (sociedade limitada)* sediada na Paraíba. Tem como missão viabilizar a transformação digital através do desenvolvimento tecnológico e de capacitações customizadas. A FABWORK é a desenvolvedora da solução NEORON uma solução ecossistema global formado por empresas que detém o selo Intel: AI Builders por desenvolverem tecnologias com IA de ponta. Para obter esse selo, a Intel aprovou a performance dos algoritmos de IA da plataforma NEORON através do uso da infraestrutura da Intel Innovation Leader / Startup revelação.

Por fim, a N2VEC é uma *startup (sociedade limitada)* sediada em Sorocaba, São Paulo. Tem como missão desenvolver soluções especializadas em buscar informações em documentos não estruturados, páginas web e texto em banco de dados. A N2VEC é a desenvolvedora da solução jus2vec, um programa de computador inteligente que faz pesquisa jurisprudencial.

A proposta está organizada da seguinte maneira:

- Etapas técnicas para a solução do problema;
- Cronograma Físico-Financeiro;
- Exemplos Relevante do Portfólio das empresas;
- Qualificação Acadêmica e Experiência Profissional;
- Metodologia.

2. Etapa Técnica para a Solução do Problema

De uma forma geral, os diversos problemas abordados durante o projeto envolvem a aplicação de técnicas distintas de Processamento de Linguagem Natural (do inglês *Natural Language Processing - NLP*), que serão explorados nos momentos propícios. Na seção a seguir vamos apresentar um breve resumo sobre essas técnicas utilizadas pela equipe do consórcio. Já na seção 2.2 abordamos como essas técnicas serão aplicadas para resolver os problemas específicos apresentados pelo projeto.

2.1. Técnicas de NLP empregadas no projeto

O projeto engloba alguns tipos diferentes de técnicas de NLP: Detecção de Entidades Nomeadas (do inglês *Named Entity Recognition - NER*), categorização automática de texto, busca e sumarização automática de texto. Nas subseção seguinte resumimos brevemente a técnica de NER e na seção 2.1.2 apresentamos a técnica de categorização automática de texto. A seção 2.1.3 descreve a técnica de busca textual.

2.1.1. Detecção de Entidades Nomeadas

Essa técnica consiste em determinar partes do texto que menciona entidades e indicar qual a categoria dessas entidades, como nomes próprios, endereços, nomes de organizações, etc. **Com essa técnica seremos capazes de extrair informações das petições para poder, entre outras tarefas, analisar a admissibilidade do processo.**

O principal desafio está em resolver certas ambiguidades, além de considerar as palavras próximas para escolher a qual categoria uma determinada entidade pertence. Considere o exemplo abaixo:

"Dr. Rodolfo S. Silva, um renomado cardiologista, adquiriu um novo consultório na Av. Edson A. Nascimento pela quantia de R$ 4.5 milhões, que atenderá pacientes do S.U.S. e do plano Saúde Mais Inc."

Nessa frase, temos as seguintes entidades:

- Rodolfo S. Silva (PESSOA)
- Av. Edson A. Nascimento (LOCALIDADE)
- S.U.S. (ORGANIZAÇÃO)
- Saúde Mais Inc. (ORGANIZAÇÃO)

Perceba que NER traz os seguintes desafios:

- Identificar qual palavra faz parte de uma entidade;
- Identificar onde começa uma entidade;
- Identificar onde termina uma entidade;
- Classificar uma entidade.

Há diversas bibliotecas para extração de entidades nomeadas disponíveis, porém, a identificação de entidades em um corpus de domínio específico ou a identificação de entidades de categorias diferentes das categorias padrões pode requerer o treinamento de um modelo de NER. Imagine, por exemplo, que deseja-se extrair a qualificação de uma pessoa mencionada no texto, como em:

"Dr. Rodolfo da Silva, **Diretor Administrativo Sênior.** do Hospital das Clínicas de São Paulo"

Nesse exemplo, o termo "Diretor Administrativo Senior" pode ser uma entidade do tipo "qualificação". Nesse caso, é necessário treinar um modelo para isso, uma vez que os modelos padrão não atendem esse tipo de categoria.

Existem diversas alternativas para o treinamento de modelos de NER para extrair entidades de categorias específicas como a do exemplo. Uma das técnicas clássicas está no uso do algoritmo CRF (*Conditional Random Field*). Trata-se de um modelo discriminativo e probabilístico que trabalha com sequências. Em outras palavras, as amostras anterior e seguinte são levadas em consideração. A expressão abaixo ilustra esse modelo:

$$P(y, X, \lambda) = \frac{1}{Z(X)} exp\{\sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y_{i-1}, y_i)\}$$

$$\text{Where: } Z(x) = \sum_{y' \in y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y'_{i-1}, y'_i)$$

No caso, *f* é o conjunto de atributos, que deve ser codificado e lambda é o conjunto de pesos aprendido pelo algoritmo. Exemplos de atributos para NER:

f1(s,i,li,li−1)=1 Se a primeira letra de li é maiúscula; senão 0

f2(s,i,li,li−1)=1 Se li é composta somente por letras maíusculas; senão 0

f3(s,i,li,li−1)=1 Se a POS Tag de li for NOUN; senão 0

Indo além da codificação manual dos atributos, pode-se também utilizar redes neurais recorrentes como a rede LSTM (*Long-Short Term Memory*), que é uma rede que aproveita o aprendizado de uma amostra anterior (no caso, de uma palavra anterior) para amostras seguintes, sendo especialmente aplicável para problemas com sequências. Esse tipo de rede é então capaz de aprender se uma entidade nomeada pode aparecer após uma determinada sequência de palavras.

As redes LSTM, ao contrário das redes neurais recorrentes convencionais, que apenas atualizam o estado aprendido em etapas anteriores, tem a vantagem de trazer o conceito de memória, que é controlada pelo modelo. Desta forma, esse modelo não apenas armazena os estados aprendidos em etapas anteriores, mas decide o que deve ser aproveitado, descartado ou atualizado. Vejamos a ilustração abaixo, extraída de http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

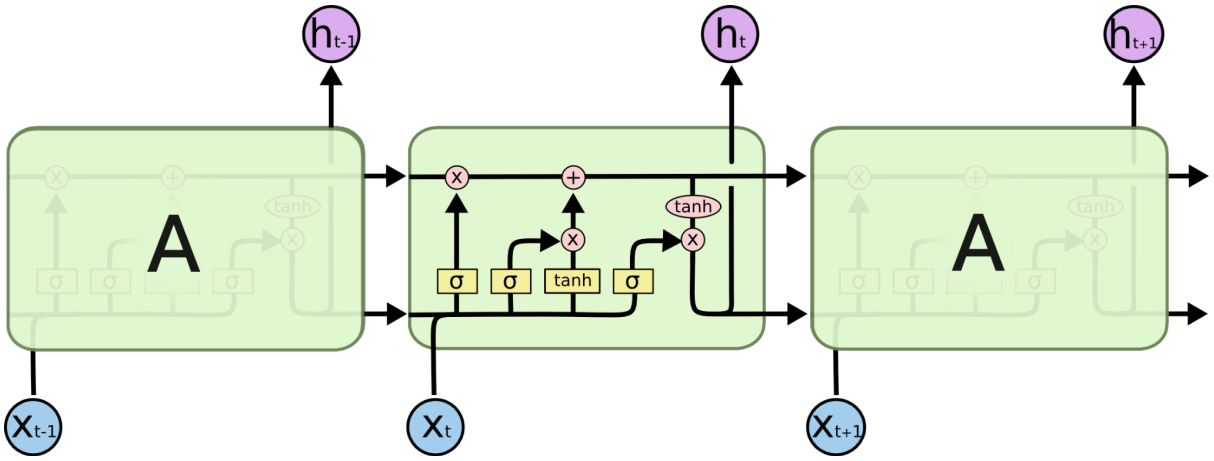


Figura 1 - Redes LSTM

As etapas ilustradas são:

1. *Qual estado de memória deve ser lembrado*: A primeira camada usa uma função sigmoid que retorna um valor entre 0 e 1 para decidir se o estado de memória vindo da iteração anterior deve ser mantido (1 para completamente mantido) ou esquecido (0 para completamente esquecido);
2. *Qual estado de memória deve ser atualizado*: Primeiramente usa-se uma função sigmoid para decidir quais valores serão atualizados com os novos dados da amostra atual, depois usa-se uma função tanh (que retorna um valor entre -1 e 1) para criar um novo vetor de estados da memória(denominado Ct) que poderá ser utilizado na iteração seguinte;
3. *Atualizar as células de memória antigas*: Combina os estados que devem ser lembrados, os estados que devem ser atualizados e os novos estados para atualizar as células de memória, que serão utilizadas na iteração seguinte;
4. *Decide a saída*: Por fim, a saída da iteração é feita através da combinação de uma função *sigmoid* e uma função tanh, considerando o estado atual da memória.

2.1.2. Categorização Automática de Texto

Outra técnica que será amplamente aplicada trata-se da categorização automática de texto. Essa técnica consiste em converter os textos em representações numéricas e aplicar algoritmos de aprendizado de máquina que aprendam a classificar os textos de

acordo com categorias pré-definidas. Como exemplo, temos a tarefa de distinguir automaticamente o tema de uma petição como de interesse público ou não. Parte dessa técnica inclui a rotulação dos dados, que pode ser feita automaticamente ou aproveitando anotações já existentes nos documentos do tribunal, ou mesmo a rotulação manual de documentos.

A primeira etapa na tarefa de categorização automática de texto consiste em realizar um pré-processamento no texto, a fim de extrair as representações numéricas que podem ser utilizadas pelos algoritmos de aprendizado de máquina. Uma das técnicas mais populares para pré-processamento de texto é a bag of words (bolsa de palavras, numa tradução livre). Esse nome é devido ao fato de que a técnica leva em conta a frequência das palavras em um texto, ignorando a ordem ou estrutura das palavras. Todavia, a representação ainda é fiel à associação entre as palavras e vem sendo utilizada com sucesso em diversas pesquisas na área.

Essa técnica consiste em gerar um vetor onde cada dimensão (ou coluna) indica a frequência de uma determinada palavra em um texto. Suponha, por exemplo, a representação de *bag of words* para as frases: “O gato subiu no telhado” e “O telhado tem goteira”. Nesse caso, haverá uma dimensão no vetor para cada palavra existente nessas duas sentenças, e as linhas representam as duas sentenças, conforme observado na Tabela 1 abaixo.

Tabela 1-Exemplo de representação de frases no modelo *bag-of-words*.

o	gato	subiu	no	telhado	tem	goteira
1	1	1	1	1	0	0
1	0	0	0	1	1	1

Embora essa representação já seja em forma de um vetor de números, ainda há o problema que algumas palavras devem aparecer com alta frequência, enquanto outras aparecem raramente, apesar de ainda serem importantes discriminantes da amostra. Para tratar esse problema, também calcula-se o TF-IDF (*Term Frequency - Inverse Document Frequency*).

Term Frequency: um termo que aparece muito em um documento, tende a ser um termo importante. Em resumo, divide-se o número de vezes em que um termo apareceu pelo maior número de vezes em que algum outro termo aparece no documento.

Inverse Document Frequency: um termo que aparece em poucos documentos pode ser um bom discriminante. Obtêm-se dividindo o número de documentos pelo número de documentos em que o termo aparece.

Na prática, os valores de TF-IDF substituem a frequência de palavras apresentada no exemplo da Tabela 1. Todavia, perceba que o modelo de *bag of words* trará uma alta dimensionalidade à amostra, uma vez que haverá tantas dimensões quanto o número de palavras total de todos os textos do corpus. Utilizar esses dados para treinar modelos de aprendizado de máquina requer uma capacidade computacional significativa, além de poder

introduzir ruídos ao modelo. Para sanar esse problema, pode-se reduzir o número de dimensões utilizando uma Decomposição em Valores Singulares (*Singular Value Decomposition* - SVD). De fato, ao utilizar o SVD, essa técnica passa a se chamar Análise Semântica Latente (*Latent Semantic Analysis* - LSA).

Outro modelo de representação muito utilizado recentemente trata-se dos vetores de palavras, ou Word2Vec. Essa técnica consiste em utilizar uma rede neural artificial simples com o objetivo de, dado uma palavra, prever a probabilidade da ocorrência de outra palavra nas proximidades. Porém, não se usa a rede neural treinada, mas os valores aprendidos como pesos da camada escondida são utilizados como representação para a palavra dada.

A Figura 2, copiada de (McCormick, 2016), mostra como são obtidas as amostras para treinar uma rede neural no modelo Word2Vec com arquitetura Skip-Gram para a frase "*The quick brown fox jumps over the lazy dog*" ("A rápida raposa marrom pula sobre o cachorro preguiçoso", numa tradução livre). Nesse exemplo, a palavra usada como entrada da rede neural é marcada em azul, enquanto o resultado esperado é uma das palavras dentro do quadro em destaque (que trata-se da janela escolhida para delinear a proximidade). As amostras utilizadas são exibidas na coluna mais à direita.

O modelo de rede neural nessa arquitetura se assemelha ao exibido na Figura 5. Nesse outro exemplo, trata-se de um conjunto de 10 mil palavras e a entrada mostrada na figura é a palavra "ants" (formigas). Observe que esta palavra é representada em forma de one hot veto-, ou seja, um vetor com 10 mil dimensões (o mesmo número de palavras), onde somente uma leva o valor 1, enquanto as demais são iguais a 0. Há uma camada de neurônios escondida e uma camada de saída. Por sua vez, na camada de saída, cada neurônio retorna a probabilidade de uma palavra estar nas proximidades (também há 10 mil neurônios, ativados pela função Softmax). Essa rede é treinada com todas as ocorrências de cada palavra no corpus, e os pesos estabelecidos para a camada escondida são escolhidos como a representação dessa palavra (nesse exemplo, um vetor de 300 elementos).

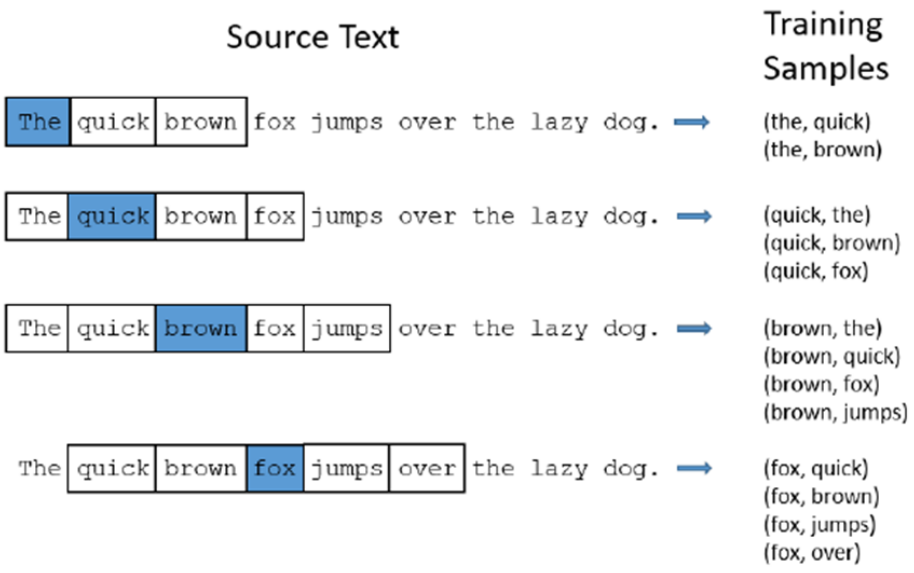


Figura 2 - Exemplo de amostra de treinamento de uma rede neural para o modelo Word2vec em inglês.

Por fim, em (Dai, 2015) foi proposto um novo modelo que estende o Word2Vec, denominado Doc2Vec. Essa técnica consiste em adicionar um vetor único ao documento (id do documento) ao conjunto de treinamento da rede neural, e é considerada adequada para representar documentos de tamanho variável. A Figura 3, retirada de (Shperber, 2017) ilustra a adição desse vetor na entrada de uma rede neural para a prever a palavra “on” (em) seguida das palavras “the cat sat” (o gato sentou) na frase em inglês “the cat sat on” (o gato sentou em…). Neste trabalho, pretendemos realizar experimentos com as diversas formas de representação no estado da arte, como BERT e suas variações, sendo uma das mais recentes o BERTikal (https://github.com/felipemaiapolo/legalnlp), desenvolvido especificamente para o domínio jurídico.

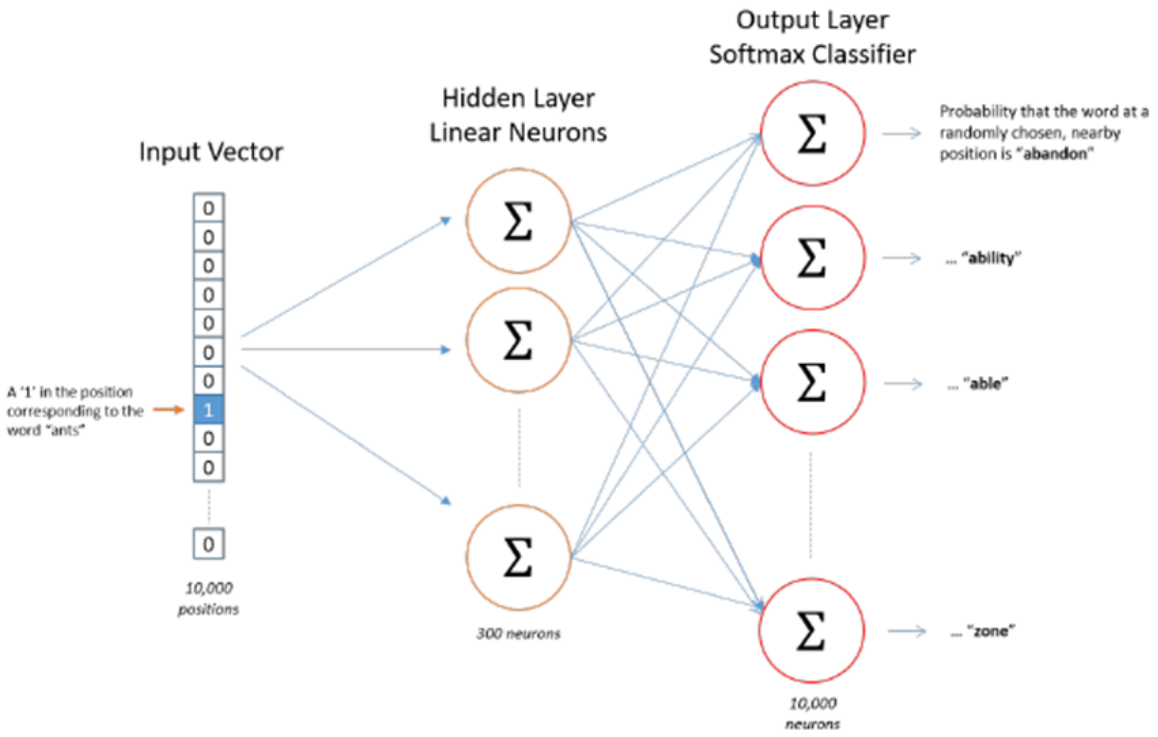


Figura 3 - Exemplo de arquitetura de rede Skip-Gram para representação de modelo Word2vec

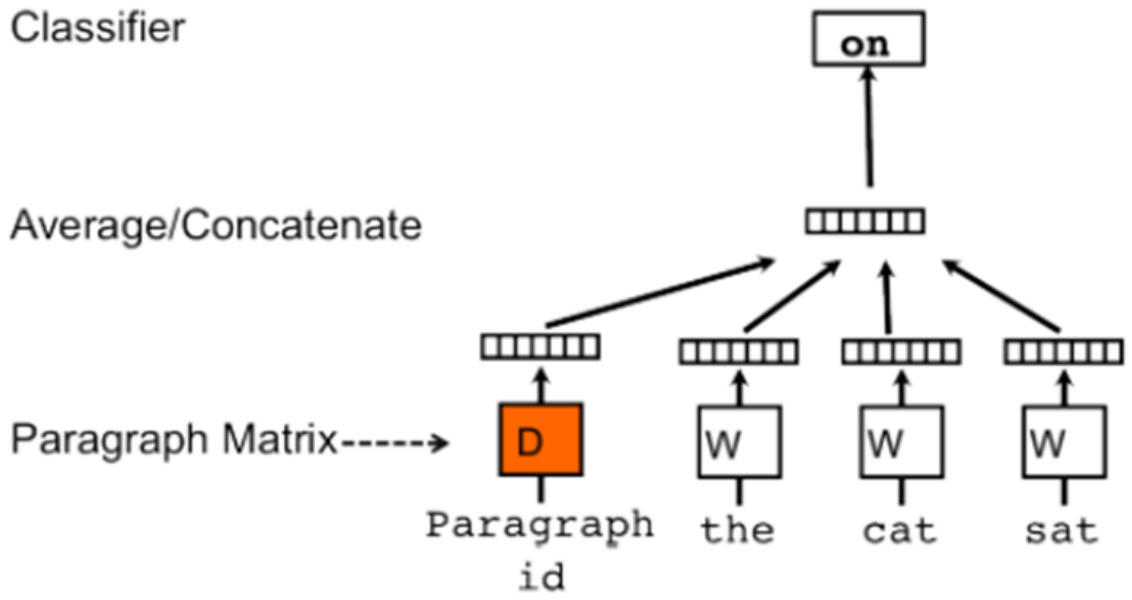


Figura 4 - Ilustração do modelo Doc2vec para prever a palavra "on" ocorrendo logo após "the cat sat" na frase "the cat sat on".

Uma vez que os textos são transformados em representações numéricas, diversos algoritmos de aprendizado supervisionado podem ser utilizados durante os experimentos diversos. Dentre os algoritmos que podem ser utilizados estão Naive Bayes, Máquinas de Vetores de Suporte (SVM) e algumas arquiteturas de redes neurais como redes convolucionais (CNN) e Multilayer Perceptron.

A experimentação deverá seguir os padrões de pesquisa comuns a projetos de aprendizado de máquina atuais. Os dados serão divididos em treino e teste (70% e 30%, respectivamente). O treino contará com seleção de hiper-parâmetros utilizando validação cruzada. Por fim, os algoritmos serão comparados usando teste de McNemmar (ou outro teste de hipótese similar que se aplique à distribuição dos resultados obtidos) para permitir escolher o algoritmo que seja significativamente melhor.

2.1.3. Busca em texto

Uma vez que dois documentos são representados como vetores numéricos, é possível comparar a similaridade entre os documentos ao calcular o cosseno do ângulo entre esses documentos. Ao comparar essa similaridade, pode-se obter uma busca textual que leva em conta a similaridade semântica das palavras.

Para calcular o cosseno do ângulo entre os documentos, representados como vetores numéricos, basta resolver a equação do produto escalar entre os vetores, para encontrar o cosseno, assim como na Figura 7 abaixo.

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Figura 5 - Calculando o cosseno do ângulo, utilizando produto escalar.

Contudo, observe que trata-se de uma medida de orientação, não magnitude. Quanto menor o ângulo entre a representação vetorial de dois documentos, maior a similaridade entre eles. A Figura 8, retirada de (PERONE, 2013) ilustra como diferentes sentenças são orientadas num espaço multidimensional.

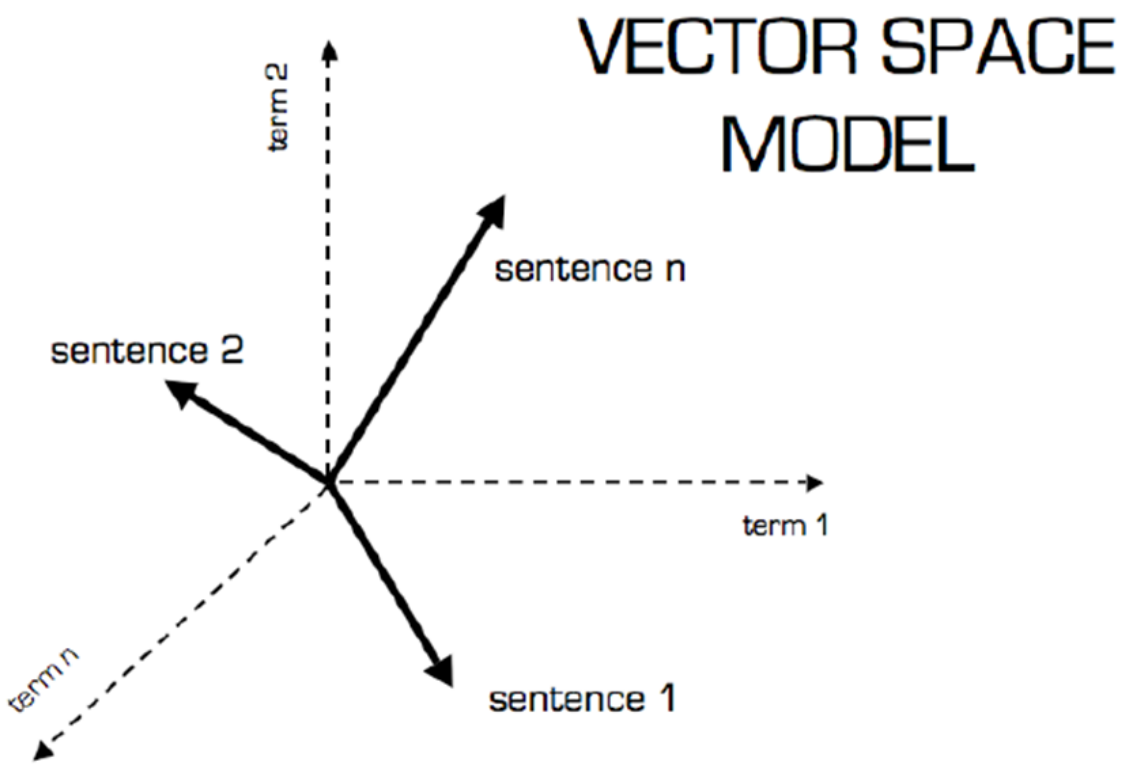


Figura 6 - Exemplo de representação de sentenças num espaço vetorial.

Neste projeto utilizaremos a similaridade de cosseno para comparar a petição com casos anteriores. Além disso, também é possível ordenar as jurisprudências encontradas com base na similaridade com a petição.

2.2. Solução para desafios técnicos

Os problemas apresentados pelo projeto envolvem a aplicação de várias das técnicas de Processamento de Linguagem Natural apresentadas na seção anterior. Em alguns casos, inclusive, é necessário combinar mais de uma técnica para a solução de um mesmo problema. Nesta seção identificamos os principais desafios técnicos e propomos uma abordagem para a solução de cada um deles.

2.2.1. Detecção de Significado em peças processuais

A detecção de significado em peças processuais dará início a partir de uma página web com formulário para upload de documento da petição inicial em formato PDF. Esse formulário será acompanhado de funcionalidades para validação de segurança antes da submissão, como captcha e cadastro de login com senha do peticionador.
Uma vez que a petição é recebida pelo formulário, o PDF será convertido em texto usando a biblioteca textract (https://textract.readthedocs.io/en/stable/) capaz de interpretar diversos formatos de arquivos, além de incluir funcionalidades para extrair texto de imagem usando OCR.

As informações da petição, como nome e endereço do autor serão extraídas utilizando bibliotecas de NER como o Spacy, a fim de preencher o formulário das Representações e Denúncias sobre Aquisições Públicas.

Já para extrair a qualificação do autor, será necessário treinar um modelo de NER conforme descrito na seção 2.1.1. Para tal, será realizada uma tarefa de anotação de qualificações mencionadas em peças anteriores para servir como base de aprendizado.

O formulário será então exibido ao usuário para confirmação das informações extraídas. A resposta do usuário (tanto positiva ou negativa) será armazenada para treinamento futuro dos modelos de NER.

Por fim, serão realizados o cálculo de probabilidade de concessão de medidas cautelares e o exame de admissibilidade, conforme descritos nas seções 2.2.2 e 2.2.3 a seguir.

2.2.2. Cálculo de Probabilidade de Concessão de Medidas Cautelares

Esta tarefa será abordada como um problema de categorização automática de texto, como descrevemos na seção 2.1.2. Todavia, antes de treinarmos nossos modelos de Inteligência Artificial capazes de discriminar a probabilidade de concessão de medidas cautelares, realizaremos um trabalho de coleta e análise de casos anteriores, e uma equipe de estagiários da área de direito será mobilizada para rotular esses documentos como havendo a necessidade de medidas ou não. Em seguida, haverá um trabalho de *Feature Engineering* para extrair informações desses documentos, preparando para ser utilizado como parâmetros para a rede neural.

Para realizar o cálculo de probabilidade de concessão de medidas cautelares, nós nos inspiramos em soluções estado-da-arte utilizadas em contexto que os dados de entrada estão estruturados de maneira semelhante e que a saída da Inteligência Artificial seja um valor probabilístico. Em nossas pesquisas, observamos soluções híbridas que atendem bem os objetivos buscados nesse trecho. Essa solução híbrida pode ser a rede neural conhecida como Wide and Deep (Mikolov, 2013), desenvolvida pela Google e que pode ser observada na Figura a seguir.

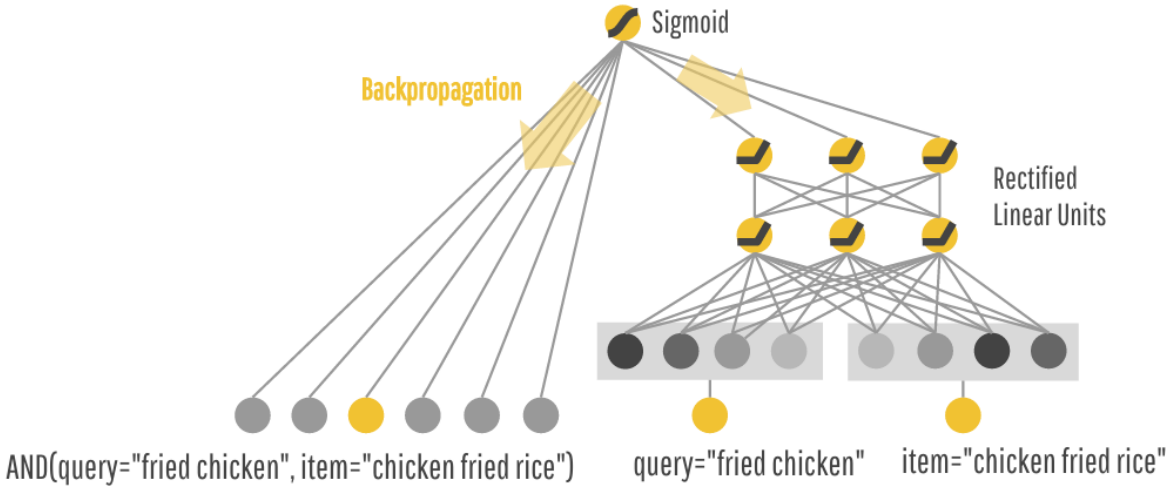


Figura 7 - ilustração da arquitetura de rede desenvolvida pela Google.

A arquitetura de rede combina dois formatos de entradas: *wide* and *deep*. No formato *wide*, poderemos ter atributos descritos, por exemplo, quantidade de vezes que uma determinada palavra-chave aparece no texto, presença/ausência de algum termo técnico que esteja associado à medida cautelar, etc. Do outro lado, o formato de dado *deep* consiste em processar um dado mais complexo, passando o dado para análise de um algoritmo com um número maior de camadas, que poderá fazer associações mais complexas, nesse aspecto, esse tipo de arquitetura de rede pode receber dados mais complexos, como o texto propriamente dito. Dessa forma, esse modelo de machine learning é capaz de analisar o dado de entrada de maneira mais aprofundada e detalhada, sendo capaz de realizar generalizações mais precisas na ausência parcial de alguma informação que pode ser complementada pelos demais atributos presentes.

Por fim, a saída desse modelo de *machine learning* é uma probabilidade. Considerando um cenário onde há dados rotulados, isto é, dados que possuem a descrição informando se a concessão da medida cautelar foi obtida (ou não), nós poderemos treinar um modelo de machine learning supervisionado utilizando como classe, a ocorrência da concessão. No caso da arquitetura de rede Wide and Deep, o modelo estimará uma probabilidade entre 0 e 100%, indicando a sua confiança que aquele documento atende os requisitos de concessão (ou seja, probabilidade máxima de 100%), variando até 0%, indicando que não há chance alguma de haver concessão da medida cautelar. O código-fonte desse algoritmo está implementado em python e publicamente disponível para uso na seguinte url: https://github.com/Lapis-Hong/wide_deep.

Além dessa arquitetura de rede, outras soluções também podem ser empregadas, como o Doc2Vec (https://radimrehurek.com/gensim/models/doc2vec.html), cuja implementação está disponível na biblioteca Gensim, também é capaz de processar textos, identificar semelhanças, e associá-los entre si, provendo uma probabilidade de que os documentos pertençam uma mesma classe. Nesse caso, essa probabilidade poderia ser utilizada para resolver o problema descrito neste ponto do projeto.

2.2.3. Exame de Admissibilidade

Um dos aspectos mais importantes em se usar machine learning é a capacidade de identificar regras de maneira automática sem você precisar necessariamente programar como identificá-las, para isso, basta que o algoritmo receba um grande volume de dados devidamente rotulados que a inteligência artificial fará o trabalho de descobrir quais regras definem as respostas que o cliente busca. Ou seja, nessa etapa do projeto, deverá ser empenhado um grande esforço no processo de rotulagem dos dados, identificando os trechos do documento, como por exemplo, há denúncia no documento? Se sim, em que trecho? Feito isso, será possível treinar um modelo de machine learning para aprender a identificar esses trechos de interesse.

A Figura seguinte representa essa mudança de paradigma. Veja que na programação tradicional é necessário que o programador forneça os dados e as regras, por exemplo, para identificar se uma peça contém denúncia ou não, é preciso que o programador defina a regra usada para identificar a existência de denúncia, o que pode não ser uma tarefa trivial e/ou objetiva. No paradigma usando *machine learning* você só precisa fornecer uma identificação, por exemplo, esse texto contém uma denúncia, e o algoritmo buscará identificar interseção de padrões entre diferentes textos e criar uma regra generalizável para diferentes contextos.

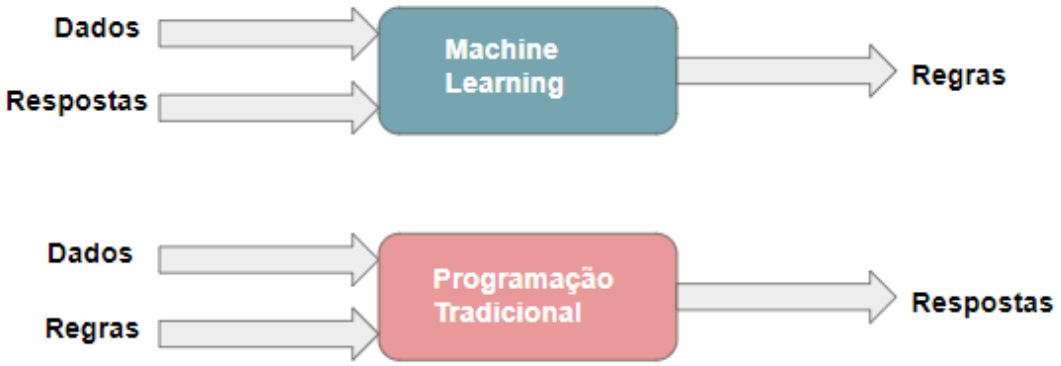


Figura 8 - Comparação entre a metodologia adotada na programação tradicional e com o uso de *machine learning.*

Com isso em mente, para atender essa tarefa, buscaremos apoio de estagiários, de preferência da área de humanas (alunos do curso de Direito, Letras, e outros), ou, utilizando

ferramentas como o Amazon Mechanical Turk[1], para realizar o processo de rotulagem de um volume de dados razoável para realização do treinamento dos modelos de *machine learning*.

Essa metodologia será utilizada para criarmos diversos modelos de inteligência artificial necessários para a análise de admissibilidade, sempre com base em casos anteriores rotulados pelos estagiários, a saber:

- Análise de Legitimidade do autor
- Análise de Redação em Linguagem Compreensível
- Análise de Indício Concernentes
- Análise de Competência do TCU
- Análise de Interesse Público

A combinação dos resultados de cada um desses modelos indicará se o processo é admissível ou não.

2.2.4. Painel de Jurimetria

O painel de jurimetria será uma interface gráfica Web acessada pelo browser a partir de um login de usuário autenticado. Essa interface mostrará as petições disponíveis na base do sistema, dando a opção de ordenar e filtrar resultados.

As informações exibidas serão as mesmas ilustradas pela figura abaixo, extraída do Termo de Referência do Edital:

Processo (veja também)	**Prioridade**	**Admissibilidade**	**Concessão de Cautelar**	**Procedência**	**Encaminhamento**
001.234 410.136 225.875	Alta	95%	78%	75%	Cautelar *inaudita altera pars*
002.456 788.991	Alta	95%	65%	Faltam provas	Diligência para saneamento
003.567 651.565 546.444	Média	90%	Não se aplica	85%	Procedência
003.789 010.014	Baixa	Interesse público não identificado	10%	18%	Arquivamento

[1] https://www.mturk.com/

Figura 9 - Exemplo de informações exibidas no Painel de Jurimetria, extraída do TR do Edital.

Para encontrar processos comparáveis, inicialmente, os processos anteriores serão indexados em representações vetoriais, em processo batch de rotina. De forma análoga, toda vez que uma petição for inserida no sistema, também será gerada uma representação vetorial da mesma, e os processos similares serão então obtidos ao realizar uma busca textual, conforme descrito na seção 2.1.3.

A probabilidade de procedência será calculada utilizando um modelo de Inteligência Artificial que será treinado utilizando os casos anteriores, que serão rotulados por uma equipe de estagiários estudantes de direito, que indicarão os procedentes e não-procedentes. As metodologias de classificação serão as mesmas mencionadas na seção 2.1.2 deste documento.

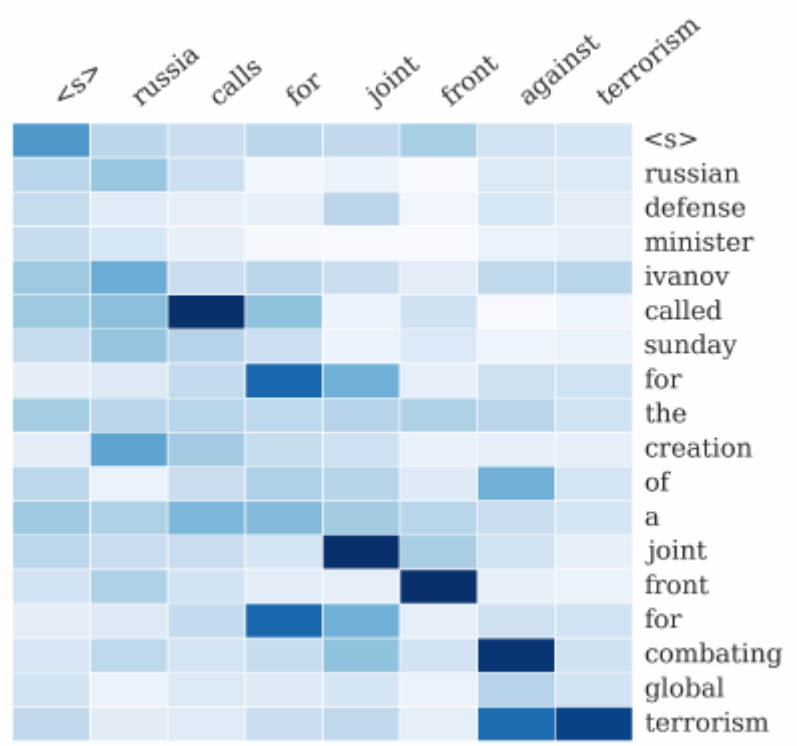
O mesmo procedimento também será realizado para criar um modelo de Inteligência Artificial capaz de indicar a prioridade do processo, também com base em casos anteriores e informações de prioridade presentes neles ou, em caso de indisponibilidade dessas informações, utilizaremos também uma equipe de estagiários estudantes de direito para rotular essas informações.

Já as informações de Admissibilidade e Concessão de Cautelar serão obtidas pelos modelos já treinados, conforme mencionados nas seções 2.2.3 e 2.2.2, respectivamente. E a proposta de encaminhamento será obtida também por modelo treinado conforme descrito adiante na seção 2.2.5.

2.2.5. Redação de Peças

O módulo de redação de peças tem início com um modelo de identificação automática da necessidade de gerar comunicação. Esse modelo de Inteligência Artificial será treinado com base em casos anteriores, também rotulados por uma equipe de estagiários estudantes de direito, e serão aplicadas as mesmas técnicas mencionadas na seção 2.1.2.

Quando houver a necessidade de comunicação, será gerado um resumo abstrato da peça. Para gerar esse resumo, utilizaremos processos anteriores e seus resumos, que serão utilizados para treinar modelos de Aprendizado Profundo (*Deep Learning*). Serão treinados dois modelos diferentes, que são estado da arte na área: o modelo NAMAS (https://github.com/facebookarchive/NAMAS), baseado em redes atencionais e o modelo PEGASUS (https://github.com/google-research/pegasus/tree/main/pegasus), baseado em um transformer que mascara parte do texto. Na figura abaixo vemos um exemplo de resultado gerado pelo NAMAS, onde o texto original aparece nas palavras ao lado direito, na vertical, e o resumo aparece no topo, na horizontal.



Por fim, será desenvolvido um modelo de Inteligência Artificial para predição de encaminhamentos. Para isso, será criada uma base de tipos de encaminhamento e os casos anteriores serão rotulados por uma equipe de estagiários estudantes de direito, que indicarão os tipos de encaminhamento de cada um deles e os excertos válidos para cada um deles. Esses dados serão utilizados para treinar modelos de Inteligência Artificial que, dado um par documento e excerto, indicará se o excerto é válido para o documento. Novamente, as técnicas serão similares ao descrito na seção 2.1.2. O encaminhamento previsto será também utilizado na exibição da petição no painel de jurimetria, conforme mostrado na seção anterior.

Haverá então uma tela de análise técnica que, quando o auditor selecionar a opção a partir do painel de jurimetria, poderá visualizar quais são os excertos sugeridos pelo algoritmo. Nessa tela, o auditor poderá indicar qual excerto é mais adequado, e também poderá indicar excertos incorretos. Esses excertos incorretos serão registrados pelo sistema, que os utilizará como amostra em novo treinamento futuro para melhorar seus resultados de forma contínua.

3. Cronograma Físico-Financeiro Proposto

No cronograma físico detalhamos a execução das tarefas compreendidas entre os prazos de início e término conforme a tabela abaixo.

Tabela - cronograma proposto para a execução das atividades.

Tarefa	Recursos	Início	Fim
Cálculo de Probabilidade de Concessão de Medidas Cautelares			
Análise de Necessidade de Medida Cautelar (Análise de Necessidade de Medida Cautelar)	**DS1**	18-ago-2023	10-out-2024
Exame de Admissibilidade			
Análise de Legitimidade do autor (Análise de Legitimidade do autor)	**DS2**	17-mai-2022	27-dez-2022
Análise de Indício Concernentes (Análise de Indício Concernentes)	**DS3**	18-mai-2022	25-abr-2023
Análise de Redação em Linguagem Compreensível (Análise de Redação em Linguagem Compreensível)	**DS2**	23-dez-2022	19-set-2023
Análise de Competência do TCU (Análise de Competência do TCU)	**DS3**	23-abr-2023	05-nov-2023
Análise de Interesse Público (Análise de Interesse Público)	**DS2**	30-out-2023	08-out-2024
Painel de Jurimetria			
Comparação com causas anteriores (Comparação com causas anteriores)	**DS4;**	18-mai-2022	12-fev-2023
Interface do Painel de Jurimetria (Interface do Painel de Jurimetria)	**FE1; BE1**	25-set-2022	05-mar-2023
Probabilidade de Precedência da Causa (Probabilidade de Precedência da Causa)	**DS4**	05-fev-2023	18-ago-2023
Priorização de Processos (Priorização de Processos)	**DS4**	10-jul-2023	05-fev-2024
Detecção de Significado em peças processuais			
Preencher automaticamente o formulário das Representações e Denúncias sobre Aquisições Públicas (Preencher automaticamente o formulário das Representações e Denúncias sobre Aquisições Públicas)	**DS1; FE1; BE1**	25-mai-2022	17-ago-2023
Redação de Peças			
Detectar necessidade de comunicação (Detectar necessidade de comunicação)	**DS5**	24-mai-2022	04-nov-2022
Instruções usando sumarização abstrativa (Instruções usando sumarização abstrativa)	**DS5**	02-nov-2022	29-jun-2023
Predição de análise técnica e propostas de encaminhamento (Predição de análise técnica e propostas de encaminhamento)	**DS5**	22-jun-2023	29-set-2024

A Tabela a seguir apresenta detalhes financeiros do orçamento proposto para execução desse projeto. A duração estimada para entrega de todos os objetivos é de **30 meses**, a coluna função identifica o cargo dos profissionais envolvidos na execução técnica das tarefas. A coluna total consiste na multiplicação das demais colunas, quantidade, a relação salário/mês, e a duração do projeto. Além disso, há outra tabela para identificação dos custos relacionados à infraestrutura. Nela identificamos os custos relacionados com assinatura do Google Colab como ambiente para armazenamento de dados e desenvolvimento do projeto, há também custos relacionados à visita técnica para melhor interação das equipes, permitir melhor acompanhamento dos progressos, alinhamento de expectativas, e direcionamento nos avanços. Além disso, há provisão de recursos para compra de equipamentos para os membros da equipe. Por fim, o nosso orçamento total consiste de **R$ 5.569.147,04** (cinco milhões, quinhentos e sessenta e nove mil, cento e quarenta e sete reais e quatro centavos).

Tabela - Orçamento estimado para execução do projeto proposto.

Recursos Humanos							
Código	**Função**	**Quantidade**	**Salário/mês**		**Duração**	**Total**	
DS1 a DS5	Cientista de Dados	5	R$	12.000,00	30	R$	1.800.000,00
FE1	Desenvolvedor de Software Front End	1	R$	8.500,00	30	R$	255.000,00
BE1	Desenvolvedor de Software Back End	1	R$	8.500,00	30	R$	255.000,00
ESTAG	Estagiários para rotulação de dados	20	R$	1.500,00	2	R$	60.000,00
OPS	Desenvolvedor de Operações (Dev Ops)	1	R$	12.000,00	30	R$	360.000,00
PM	Gerente de Projeto	1	R$	12.000,00	30	R$	360.000,00
PO	Dono do Produto	1	R$	12.000,00	30	R$	360.000,00
SM	Scrum Master	1	R$	8.500,00	30	R$	255.000,00
						R$	**3.705.000,00**

Infraestrutura e outros custos							
	Despesa	**Quantidade**	**Valor**		**Duração**	**Total**	
	Assinatura Google Colab Pro+	5	R$	258,00	30	R$	38.700,00
	Notebooks de trabalho	8	R$	4.500,00		R$	36.000,00
	Instâncias de VM em Google Cloud	4	R$	585,51	30	R$	70.261,20
	Despesas com viagens	10	R$	3.000,00		R$	30.000,00
	Despesas Operacional e Administrativas	1	R$	10.000,00	20	R$	200.000,00
						R$	**374.961,20**

Custo Fixo do Projeto	**R$**	**4.079.961,20**
Remuneração Variável	**R$**	**1.223.988,36**
Impostos Estaduais e Municipais	**R$**	**265.197,48**
Valor Total do Projeto	**R$**	**5.569.147,04**

Haverá um marco do projeto a cada mês, sendo composto pelo andamento de 2 sprints (de 15 dias), totalizando 30 marcos no projeto. Na tabela abaixo mostramos os marcos e seus respectivos valores. As funcionalidades entregáveis de cada marco serão delineadas pelo andamento previsto em cronograma. Para as faixas de lucro da remuneração variável adicional (além da mencionada na tabela abaixo), propomos o lucro adicional de 1,15% para cada 1% de acurácia melhorada nos modelos de inteligência artificial apresentados originalmente.

Marco	Data	Custo Fixo	Remuneração Variável	Impostos	Valor Total
1	01/06/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
2	01/07/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
3	01/08/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
4	01/09/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
5	01/10/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
6	01/11/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
7	01/12/2022	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
8	01/01/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
9	01/02/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
10	01/03/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
11	01/04/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
12	01/05/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
13	01/06/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
14	01/07/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
15	01/08/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23

Marco	Data	Custo Fixo	Remuneração Variável	Impostos	Valor Total
16	01/09/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
17	01/10/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
18	01/11/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
19	01/12/2023	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
20	01/01/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
21	01/02/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
22	01/03/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
23	01/04/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
24	01/05/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
25	01/06/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
26	01/07/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
27	01/08/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
28	01/09/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
29	01/10/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
30	01/11/2024	R$ 135,998.71	R$ 40,799.61	R$ 8,839.92	R$ 185,638.23
					R$ 5,569,147.04

4. Exemplos Relevantes do Portfólio das Empresas

4.1. FabWork

Com objetivo de apresentar justificativa dos requisitos de qualificação técnica da Fabwork, listamos alguns exemplos de projetos completamente desenvolvidos pela empresa.

4.1.1. Experiência comprovada em projetos de *Big Data Analytics* e desenvolvimentos de artefatos em Ciência de Dados

No projeto que executamos para o **SEBRAE PARAÍBA**, a FABWORK ficou responsável por acessar as bases de dados disponíveis sobre os cursos realizados pelo **SEBRAE PARAÍBA** ao longo de um período e identificar, mediante modelos de ciência de dados, personas (ou seja, *cluster*/segmentos de pessoas consumidoras) que consomem estes produtos visando direcionar melhor o planejamento de execução e comunicação da entidade. Neste projeto em particular, desenvolvemos artefatos em ciência de dados em Jupyter Notebooks para demonstrar o processo de segmentação/clusterização relacionado ao agrupamento de objetos que são semelhantes entre si e diferentes dos objetos pertencentes a outros clusters. O processo de clusterização utilizado para atendimento dos requisitos deste projeto foi o K-*medoids*, pois diferentemente de outras formas de

clusterização como K-*means* e x-*means* (que definem como o centro do cluster pontos imaginários baseados em médias) esse projeto necessitou que o centro do *cluster* fosse retirado do conjunto de dados fornecido pelo cliente.

A FABWORK ainda liderou uma consultoria com aplicação de algoritmos de aprendizagem de máquina com foco na aplicação de modelos de predição para o **CARREFOUR**, projeto direcionado para o Carrefour Express - sendo aplicado uma série de stacks de de ciência de dados em Python (Pandas, Scikit Learn, Folium, Numpy e Numba). A FABWORK ficou responsável pela camada de mineração de dados e preparar as bases de dados enviadas pelo cliente, bem como processar modelos utilizando técnicas da ciência de dados propondo um novo modelo capaz de direcionar pontos viáveis e rentáveis de abertura de nova loja do Carrefour Express com satisfatório grau de acurácia. Para a geração do mapa representativo gráfico das recomendações de pontos de novas lojas do Express, rodamos dentro do Jupyter Notebooks.

Além do mais, visando conduzir uma Prova de Conceito para o cliente **GERDAU**, cujo escopo era realizar e documentar as análises de três ferramentas de ciência de dados com (INFOSYS NIA, RAPIDMINER 9.2 e KNIME 3.7), denominadas ferramentas de Auto-ML, ou seja, auto machine learning (aprendizagem de máquina automática), a FABWORK ficou responsável por apresentar os resultados de diagnóstico para o time de ciência de dados da GERDAU, destacando ao final do relatório, recomendação de qual ferramenta estaria mais aderente aos critérios técnicos sugeridos FABWORK, como (1) suporte da ferramenta com cloud computing exemplo s3 e google cloud; (2) integração com outras tecnologias de Big Data, exemplo Hadoop e Spark; (3) suporte a banco de dados NoSQL exemplo Mongodb Cassandra; (4) suporte a gráficos, como análise de séries temporais, dentre outros.

4.1.2. Experiência com projetos de consultoria com serviço de manipulação/conexão de banco de dados (SQL - *Standard Query Language Server*)

No Projeto Data River API para o cliente **UR COMPANY**, o uso da linguagem SQL (Standard Query Language) foi aplicado na construção do banco de dados. A finalidade principal foi dar suporte a um sistema de CRUD (Create, Read, Update, Delete), voltado para as bases e fontes de dados públicas (principais entidades tratadas no sistema) relacionadas ao projeto. Além disso, foi aplicado também na construção um sistema para o registro de logs, o qual é utilizado para registrar as consultas realizadas nas bases e fontes cadastradas no Projeto Data River.

Para a construção de toda a infraestrutura do banco foi criado inicialmente um Modelo Relacional, onde foram especificados todos os relacionamentos entre as tabelas do banco e também as informações que cada uma têm como responsabilidade armazenar. Feito isso, foi escolhido o SGBD (Sistema de Gerenciamento de Banco de Dados), esse utilizado para a realização do gerenciamento do banco. O escolhido foi o PostgreSQL,

devido aos recursos disponíveis e por ser open-source e ter uma comunidade ativa que garante a continuidade e suporte do mesmo.

Na implementação e conexão do banco com a aplicação foi utilizado um ORM (mapeamento objeto-relacional), uma técnica de desenvolvimento utilizada para reduzir a contraposição da programação orientada a objetos utilizando bancos de dados relacionais. O framework ORM escolhido para isso foi o JPA (Java Persistence API), visto que esse já vinha por padrão em uma das tecnologias solicitadas pelo cliente. O framework JPA diminui o contato direto com query's SQL cruas, tornando a interação com o banco mais abstrata a nível de programação, onde só em alguns casos há a necessidade real de criar comandos SQL de forma 'crua' para buscar ou tratar alguma informação no banco. Sendo assim, a utilização da Linguagem SQL se deu da forma descrita acima.

4.1.3. Experiência com projetos de consultoria com serviço de análise e localização de bases de dados, criação e mescla dinâmica, além de importação de dados utilizando os recursos

A FABWORK executou para o cliente **UR COMPANY** os seguintes processos: (1) análise e localização de fontes de dados, visando a construção de uma base de dados para referenciar as fontes, (2) criação e mescla dinâmica de diversas fontes de dados coletadas de bases públicas disponíveis na web, como (i) modo 1 - integração via API (do inglês Application Programming Interface), (ii) modo 2 - uso de web crawling/scraping e (iii) modo 3 - abertura de página web pré-mapeada + plugin. Como processo final do projeto, a FABWORK integrou esses dados produzindo um data lake confiável e acessível para usuários em uma visão de front-end. A FABWORK fez uso de Spring Boot, FastAP! e Django REST Framework.

4.1.4. Experiência com projetos de consultoria com serviço de gerenciamento de infraestrutura em computação em nuvem - *cloud computing*

Os trabalhos desenvolvidos pela FABWORK para fins internos e externos, são documentados e tem registros disponíveis em ambiente virtual compartilhado pelos participantes, bem como há uso efetivo de infraestrutura em computação em nuvem, como ambientes do GCP (*Google Cloud Platform*) e AWS (*Amazon Web Services*). Dentro de tais ambientes, destacamos o desenvolvimento de arquitetura e infraestrutura utilizando a metodologia serverless, como o Kubernetes para gerenciamento e escalabilidade de contêineres do Docker, e também de funções distribuídas na cloud, por meio do AWS *Lambda* e do GCP *Cloud Functions*. Também executamos a gerência e monitoramento de recursos de computação e balanceamento de carga de clusters de instâncias do Compute Engine e de máquinas virtuais do AWS EC2.

4.1.5. Portfólio de cursos de formação profissional e organizacional em ciência de dados, mineração de dados e *data analytics*

A FABWORK já executou a formação de 8 (seis) turmas da academia "*Data Science* para Não Programadores", somando um total de quase 400 profissionais capacitados. Mesmo sem a necessidade de conhecimento prévio em programação, os profissionais capacitados pela FABWORK aprenderam a aplicação de ciência de dados e extrair potencial dos dados através de data analytics com foco nos seus diferenciais competitivos. Toda a academia "*Data Science* para Não Programadores" da FABWORK utiliza RAPIDMINER 9.2 ou KNIME 3.7, sendo ambas as ferramentas destaques na lista da Gartner. Das 6 turmas, 1 (uma) foi entregue no modelo in company recentemente para 30 colaboradores da **LEROY MERLIN** - rede multinacional de lojas de origem francesa de materiais de construção, decoração, jardinagem e bricolagem.

Além de turmas exclusivas da academia de "Data Science para Não Programadores", a FABWORK elaborou o programa corporativo "Programa *Data Driven*" para a **GERDAU**, denominado Programa G.Data. Foram avaliados 250 participantes, sendo assim selecionados 25 participantes que receberam 56 horas de conteúdo de capacitação on-line e sessões de mentoria técnica individual. A capacitação foi estruturada a partir da exploração de módulos de conhecimentos evolutivos para consolidar a formação evolutiva dos colaboradores dentro do programa, como (1) Introdução Geral à Ciência de Dados; (2) Lógica de Programação e Noções de Python; (3) Manipulação e Visualização de Dados; (4) Análise Exploratória de Dados; e, (5) Introdução ao Aprendizado de Máquina. Para solidificar a transformação do conhecimento em aprendizado prático pelos 25 participantes, formatamos momentos com sessões de mentoria técnica entre cada participante e a equipe da FABWORK para discutir oportunidades e formatar projetos ou ações em suas áreas e atuação que gerem valor.

4.1.6. Experiência com projetos de consultoria com aplicação de modelos de aprendizagem de máquina e processamento de linguagem natural (NLP - *Natural Language Processing*)

A FABWORK é a empresa de tecnologia responsável pelo desenvolvimento da plataforma como serviço (PaaS - *Platform As a Service*): NEORON - plataforma de construção, treinamento, gestão e integração de chatbots (robô conversacional), que permite ao usuário sem conhecimento de programação construir a experiência do seu cliente com IA conversacional, maximizando benefícios e minimizam os esforços no processo de jornada do cliente (customer experience - CX).

A tecnologia da NEORON engloba aprendizagem de máquina voltada ao NLP para interagir com o usuário com precisão e acurácia significativa, mediante aplicação de uma série de algoritmos proprietários, além de permitir a captura de dados de forma rápida, conveniente e simplificada, sem a necessidade de longos formulários. A NEORON permite a integração com canais de comunicação, como páginas customizadas (Landing Pages) e bolhas de conversação integradas ao site (Widget).

O modelo proprietário de IA-NLP da NEORON pode ser explicado a partir da execução de 8 processos:

1. NEORON recebe um texto (input);
2. NEORON envia o texto para algoritmos de pré-processamento;
3. NEORON busca por candidatos à resposta em seu treinamento;
4. NEORON compara os resultados usando aprendizagem de máquina;
5. NEORON gera um nível de confiança para cada comparação;
6. NEORON seleciona a comparação com maior nível de confiança;
7. NEORON escolhe a resposta desejada, e
8. NEORON recupera a resposta do database para exibição (*output*).

A FABWORK recebeu uma demanda da empresa GRUPO LIFE BRASIL, especialista em seguros de vida no Brasil, fruto da operação da antiga Mitraseg Corretora de Seguros. A empresa está avançando com o seu modelo de negócios de franquias denominado "*Life Brasil Franchising*", que tem o objetivo de abrir franquias por todo o território brasileiro. Para começar um projeto piloto com a empresa Grupo Life Brasil, a FABWORK foi convidada para criar um chatbot em uma página customizada (*Landing Page*) guiando as opções de comunicação automatizada para captura de dados através das conversações.

Um outro chatbot criado pela FABWORK foi destinado para a empresa Soul Consciente, que buscava consolidar um posicionamento cada vez mais digital, além do comprometimento com a inovação para transformar a maneira de se relacionar com o seu cliente, mais especificamente o setor de relacionamento com os "clientes-parceiros". Este chatbot está em formato *landing page*, disponibilizado pela NEORON.


CERTIFICADO de APOIO

COVID ZERO

Certificamos e agradecemos

FABWORK SERVICOS DE CONSULTORIA EM GESTAO EMPRESARIAL LTDA

CNPJ: 30.869.179/0001-89

por apoiar o projeto social COVIDZERO com a sua tecnologia NEORON entre Maio e Setembro de 2020 para ajudar a população brasileira durante a pandemia da COVID-19 com sintomas da doença.

16 de Outubro de 2020

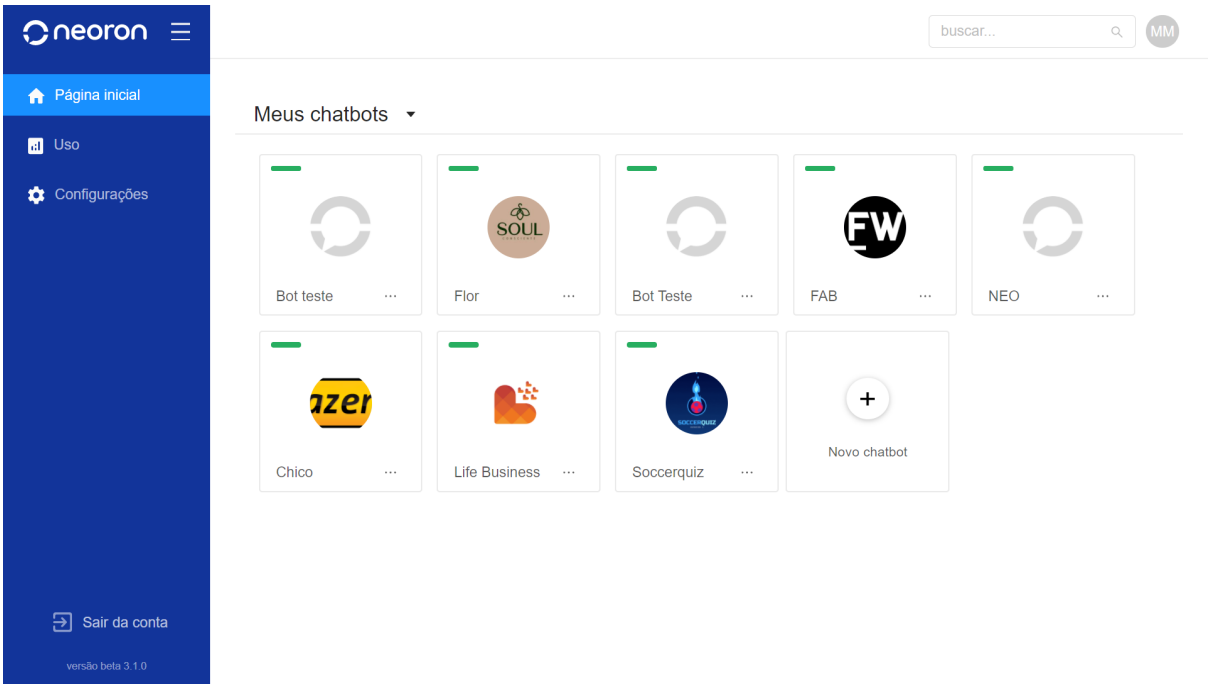
PABLO KLEIN
Presidente

covidzero.com.br

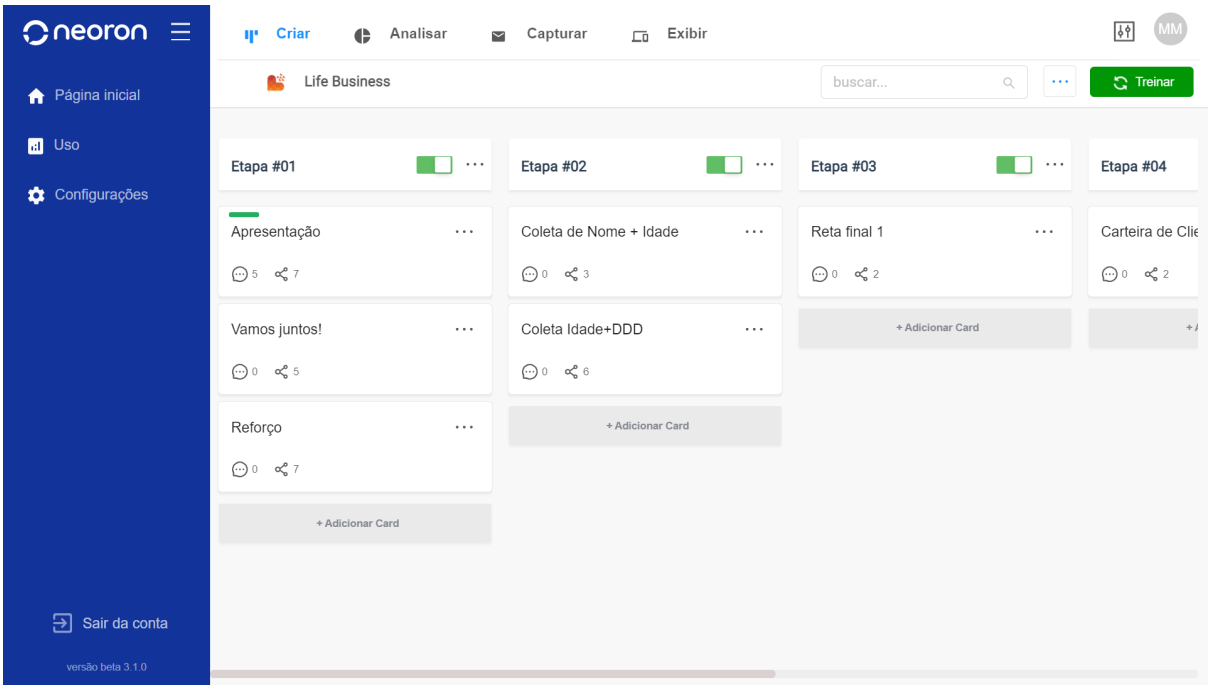
Através de um chatbot criado e gerenciado através da plataforma NEORON, a FABWORK apoiou o Projeto CovidZero, que foi liderado por um conjunto de profissionais espalhados pelo Brasil. O projeto contava com o apoio técnico de empresas de grande porte no ramo da tecnologia, como Google Brasil e Cielo. A missão do chatbot da NEORON no Projeto CovidZero foi auxiliar na triagem inicial dos sintomas do Covid-19 através de chatbot e integrar os dados a um sistema de telemedicina, sendo a partir disso um processo onde os usuários pudessem de fato apresentar os sintomas do Covid-19 para que fossem, de acordo com o diagnóstico, direcionados para provável consulta gratuitamente por médicos voluntários.

Destacamos abaixo as funcionalidades da NEORON, assim apresentando um conjunto de imagens de sua tela e detalhando todo o método NEORON para criação e desenvolvimento de chatbots.

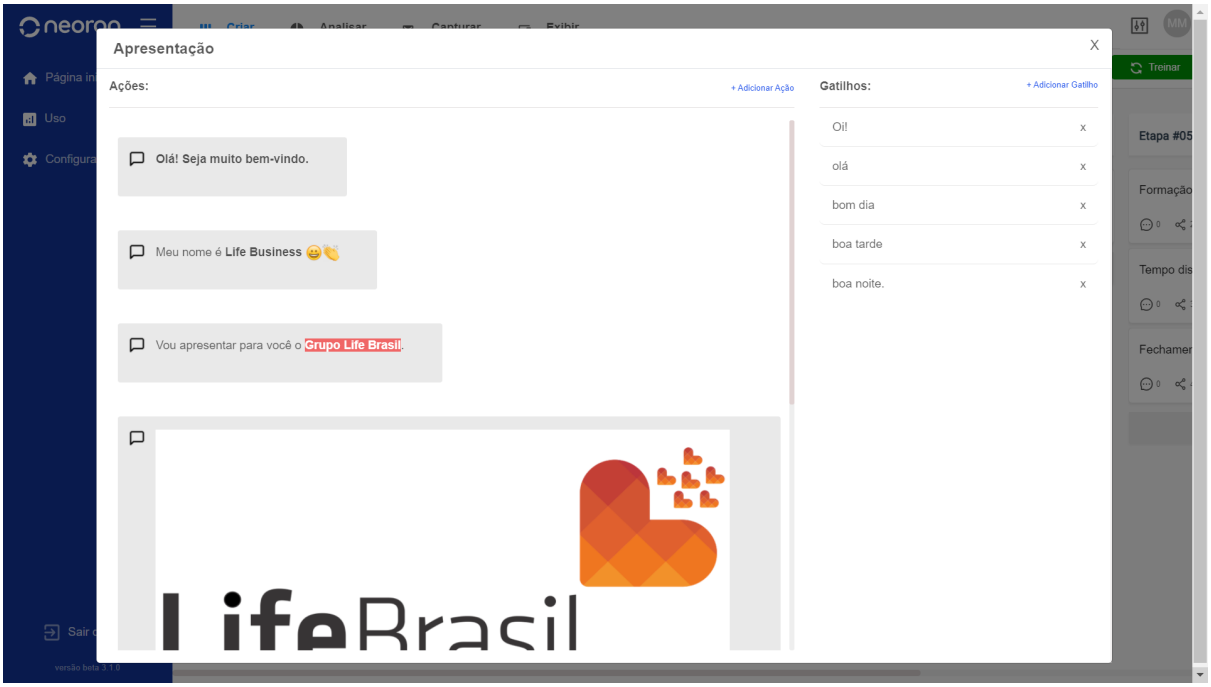
A) **Gerência de projetos de chatbot:** através da interface da NEORON, é possível gerenciar os chatbot com simplicidade e facilidade em uma única tela:



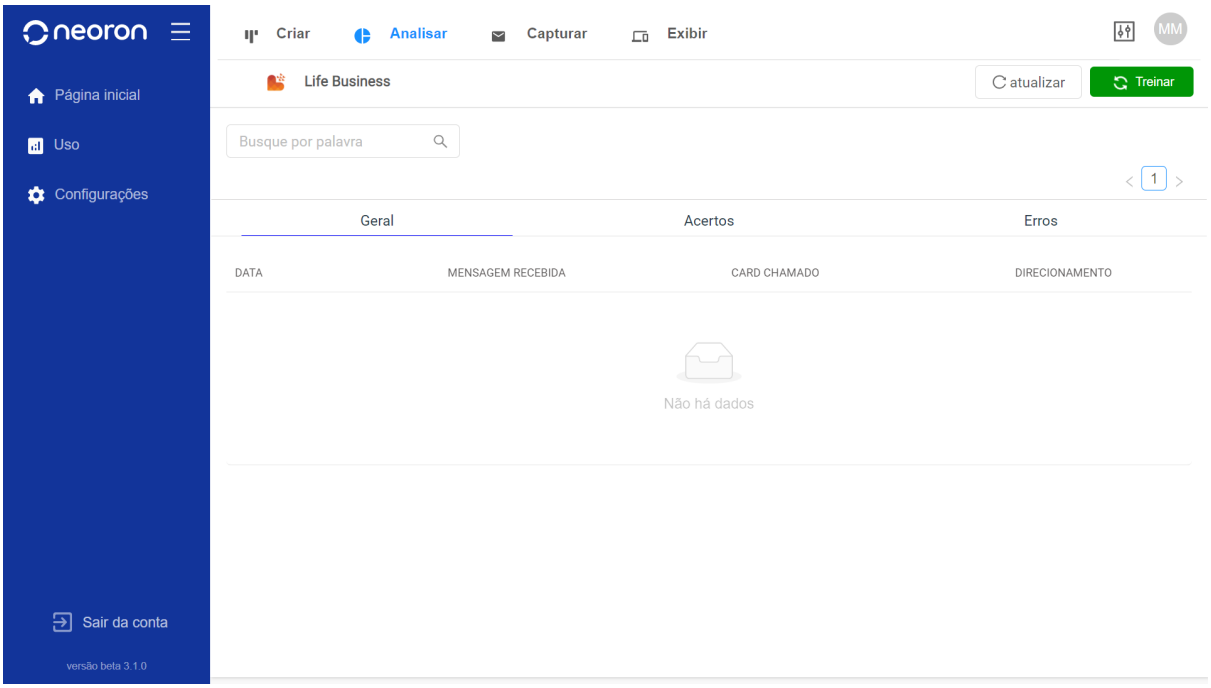
B) Criar: ambiente para desenvolvimento do chatbot baseado na estratégia de comunicação da organização, utilizando criador e gestor da base de conhecimento e fluxo conversacional, sendo inspirado no método Kanban, no qual o chatbot é criado utilizando Cards que representam tanto o fluxo conversacional, quanto à base de conhecimento.



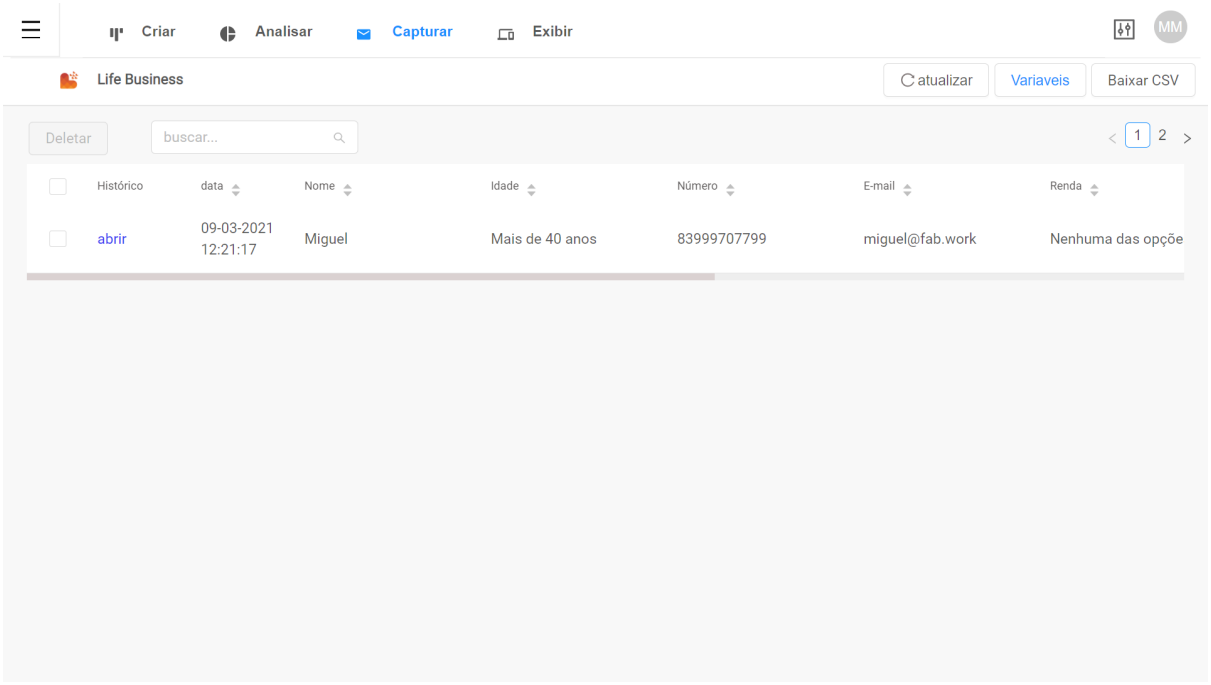
Segue imagem dentro de um Card da NEORON:



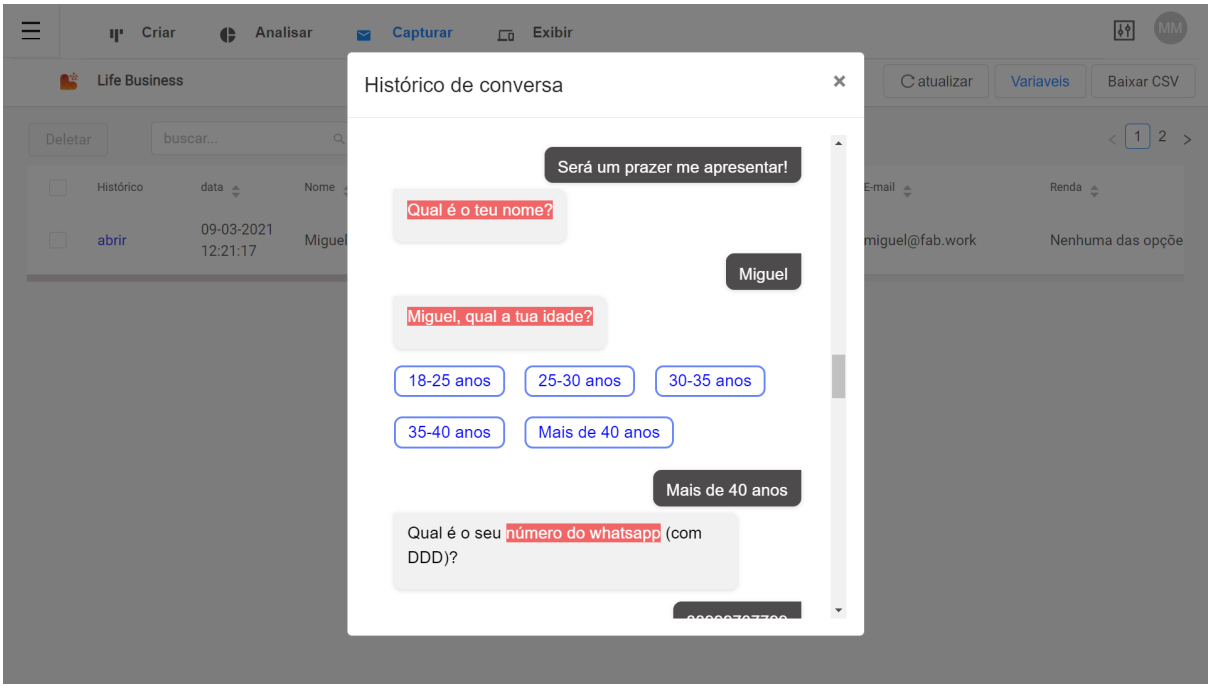
C) Analisar: ambiente de curadoria para melhoria contínua do chatbot. Aqui é permitido a melhoria do chatbot via monitoramento de atividades executadas, supervisionando a sua aprendizagem.



D) Capturar: através do ambiente de captura de dados frutos das conversações do chatbot, é possível analisar os dados capturados pelo chatbot, além de exportá-los em CSV, para que possam ser utilizados para diversos fins.

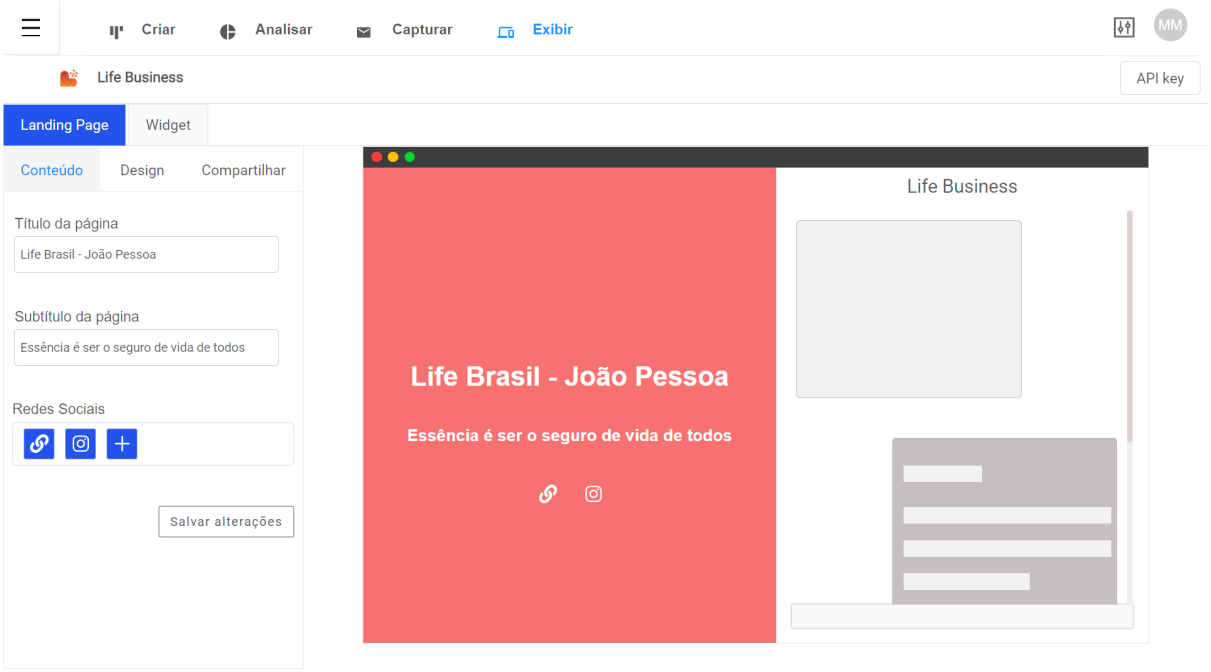


Segue imagem do log da conversação completa na NEORON:

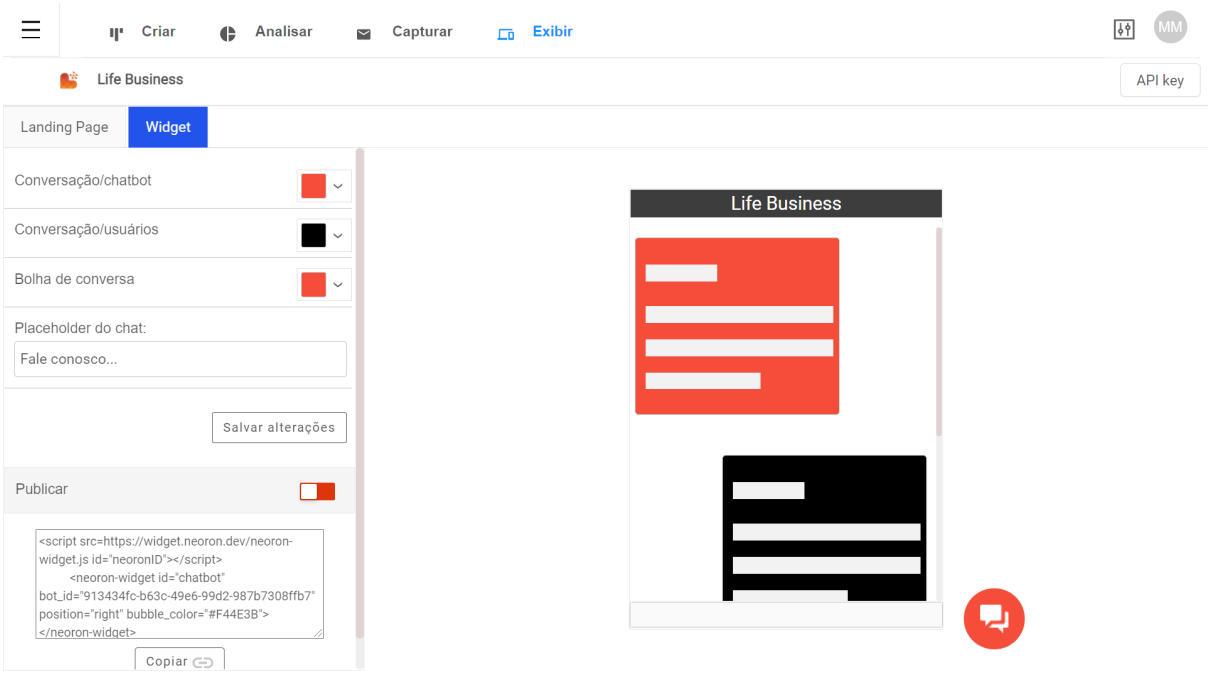


E) Exibir: o chatbot da NEORON pode ser implantado e disponibilizado para exibição pública em canais web, como a Landing Page ou integrado através da funcionalidade de Widget, além do WhatsApp.

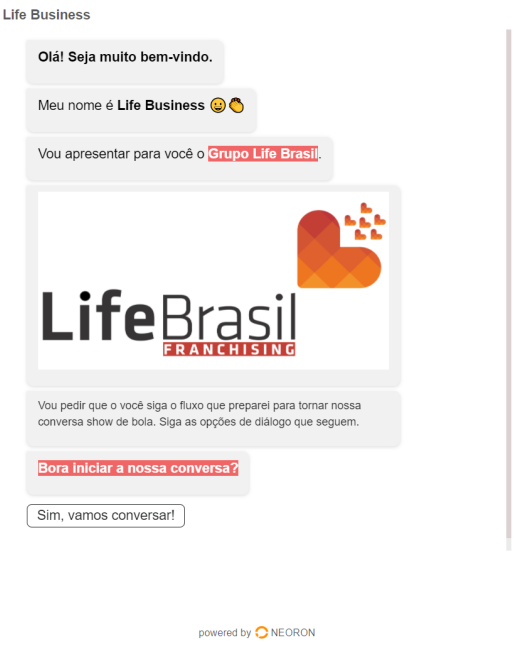
Segue imagem da customização da *Landing Page* (página customizada) na NEORON:



Segue imagem da customização da funcionalidade de Widget na NEORON:



F) NEORON.chat: interface disponibilizada para exibição pública exibição de chatbot *Landing Page* (hospedada no domínio neoron.chat ou domínio customizado), conforme interesse do cliente.

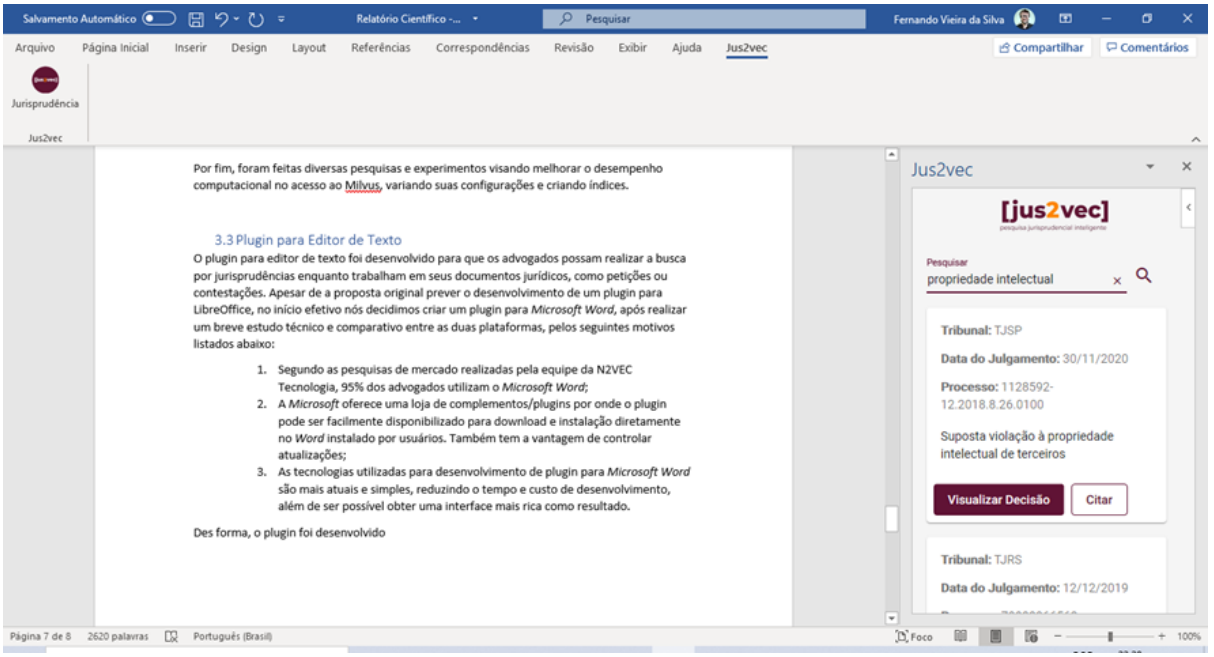


Perante o exposto acima, fica evidente a experiência técnica, os conhecimentos práticos e a garantia dos Requisitos de Qualificação Técnica da FABWORK SERVIÇOS DE CONSULTORIA EM GESTAO EMPRESARIAL LTDA –, sendo assim distribuídos entre os nossos projetos de consultorias, projetos de desenvolvimento tecnológico de produtos analíticos e digitais, especialmente quanto ao nosso domínio e propriedade da NEORON, além dos serviços prestados de capacitação e mentoria técnica.

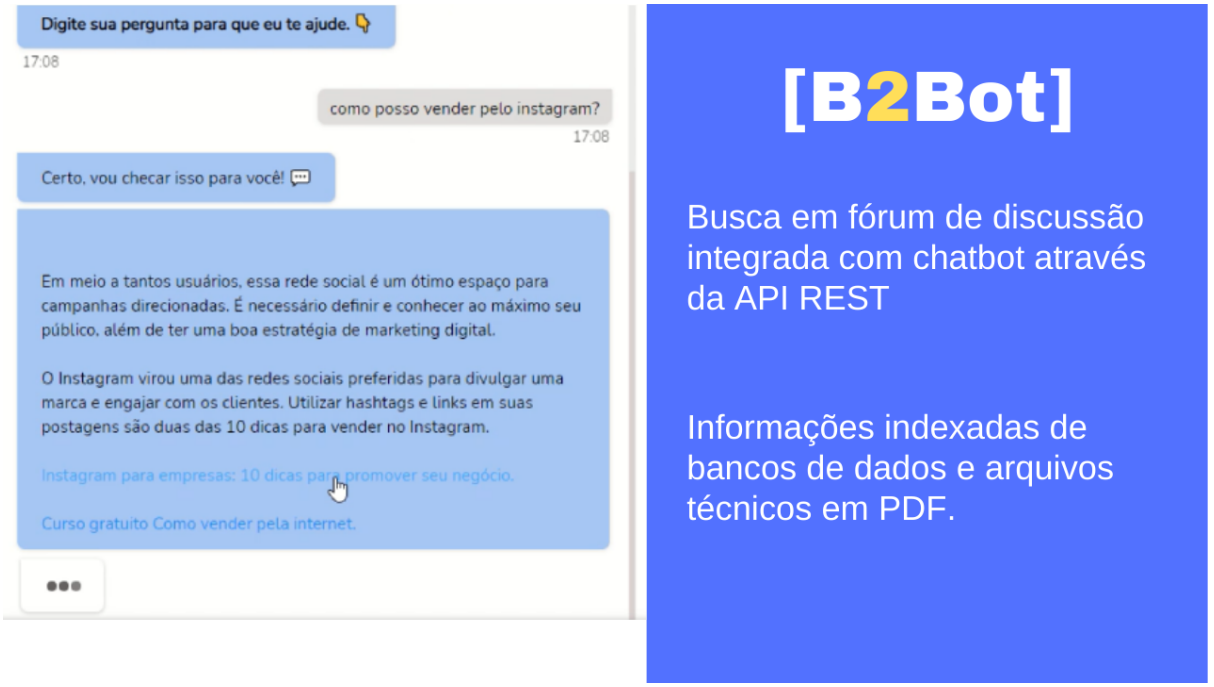
4.2. N2VEC

A N2VEC possui experiência no desenvolvimento de sistemas jurídicos utilizando Inteligência Artificial, ao desenvolver o projeto Jus2vec[2], a primeira e única ferramenta de pesquisa jurisprudencial no Brasil a ser integrada ao MS Word. O Jus2vec conta com uma engine de busca inteligente utilizando as mais modernas técnicas de Aprendizado Profundo. Essa tecnologia foi desenvolvida e vem sendo aprimorada com financiamento do programa PIPE da FAPESP.

[2] https://youtu.be/r9WWd5WpOSQ



Além do Jus2vec, a N2VEC também possui experiência em busca inteligente em outros cenários, sendo finalista da chamada pública para Inovação Tecnológica para o desenvolvimento de Serviço Digital Automatizado de Respostas Técnicas do SEBRAE.



A N2VEC participou do programa Startup SP do Sebrae e do programa de internacionalização de startups no Canada, organizado pela Dream2b e pelo Spark Centre.

4.3. RBCIP

A Rede Brasileira de Certificação, Pesquisa e Inovação desenvolve trabalhos por meio de ações formuladas pelo seu observatório, conselho de inovação e por meio dos membros associados. Os seguintes projetos foram desenvolvidos e são relevantes para apresentação da capacidade técnica da entidade:

- **Laboratório de Apoio à Inovação da Educação Básica do Brasil (LabInova)**
    - Projeto desenvolvido em parceria com Fundação de Apoio à Pesquisa, ao Ensino e à Cultura (Fapec)
    - Objetivo do programa: Fortalecer e apoiar a UFMS no processo de ampliação e ensino-aprendizagem com base no tripé Educação, Tecnologia e Inovação
    - Ações:
        - Mapeamento das ferramentas e recursos tecnológicos.
        - Desenvolvimento de modelo de disseminação da cultura da inovação;
        - Operacionalizar o fluxo e controle dos seminários regionais (online);
        - Relatórios pedagógicos por meio de bases de dados php, tratamento de dados;
        - Gerar o relatório Nível de Serviço Comparado
- **Caminhões da Tecnologia by Mobtech.**
    - Projeto desenvolvido em parceria com a Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF)
    - Objetivo: Elaboração, implementação e avaliação de projeto de pesquisa e/ou plano de trabalho para idealização, implementação e gestão de programa de capacitação em estrutura inerente para promoção da inclusão e da conectividade digital por meio da oferta de cursos de robótica, programação e novas tecnologias em zonas periféricas e de baixa renda no Distrito Federal.
    - Ações:
        - Painel automático de levantamento de dados automatizada;
        - Gestão de ativos logísticos;
        - Gestão educacional;
        - Leitura de formulários em OCR;
        - Treinamento em python, robótica e programação;
        - Estruturação dos dados em painel;
        - Coleta de dados via formulário de inscrição, análise dos dados para validação via linguagem Python.
        - Realização de pesquisas por meio de técnicas econométricas;
- **Outras iniciativas**
    - Desenvolvimento de Painel de dados inteligentes - FAPEC;
    - Colaboração na elaboração do projeto de Fundo de Capital de Risco para apoiar empresas do Parque Tecnológico do Distrito Federal
    - Estudo de governança para Núcleos tecnológicos;
    - Promoção de ecossistemas rurais de inovação por meio do Conselho de Inovação. Objetivos: promover por meio da computação em nuvem, os mecanismos de eficiência para a tomada de decisão e outros.

5. Qualificação Acadêmica e Experiência Profissional dos Principais Envolvidos

5.1. Fabwork

FABWORK é um centro de inovação de impacto e transformação digital. Desde 2019, a FABWORK é membro Intel AI Builders[3]: ecossistema da Intel Corporation para acelerar o desenvolvimento e a adoção de tecnologias com Inteligência Artificial (IA). Estamos estruturados em três verticais de atuação, as quais são:

1. **FAB TECH:** Vertical de desenvolvimento de soluções exponenciais e customizáveis de tecnologia com foco em soluções com Inteligência Artificial, *Big Data Analytics* e Internet das Coisas (do inglês, *Internet of Things* - IoT).
2. **FAB CORPORATE:** Vertical de desenvolvimento de programas de educação corporativa e mentoria técnica com foco em transformação digital, inovação de impacto e ciência de dados.
3. **FAB ACADEMY:** Vertical com um conjunto de academias (isto é, cursos rápidos e práticos) com foco no desenvolvimento de *soft/hard skills* para formar profissionais exponenciais (Profissional 4.0).

A Fabwork é composta por diversos profissionais com competências notáveis, entre eles, vamos destacar alguns desses colaboradores:

5.1.1. Edilson Ferneda

Graduado em Tecnologia de Computação pelo Instituto Tecnológico de Aeronáutica - ITA (1979), Mestre em Sistemas e Computação pela Universidade Federal da Paraíba – UFPB [hoje Universidade Federal de Campina Grande - UFCG] (1988) e Doutor em Ciência da Computação pelo *Laboratoire d'informatique, de robotique et de microélectronique de Montpellier* - LIRMM/CNRS, França (1992), com ênfase em Aprendizagem de Máquina. Entre 1986 e 2004, foi professor do Departamento de Sistemas e Computação da Universidade Federal da Paraíba [hoje UFCG], tendo atuado nos cursos de Bacharelado em Ciência da Computação, Mestrado em Informática e Doutorado em Engenharia Elétrica. Desde 2001 é professor titular da Universidade Católica de Brasília, onde atua no Curso de Bacharelado em Ciência da Computação e no Mestrado em Governança, Tecnologia e Inovação. Leciona Inteligência Artificial na graduação e, no Mestrado, Ciência de Dados. Tem experiência como consultor de organismos internacionais em projetos junto a órgãos do Governo Federal Brasileiro. No setor privado, junto à empresa FabWork, desde 2020, vem dando mentoria em projetos de Ciência de Dados nas empresas Gerdau e Leroy Merlin. Atua como avaliador de trabalhos científicos em eventos como *International Conference on Enterprise Information Systems* (ICEIS) e *International Conference on*

[3] https://builders.intel.com/ai

Agents and Artificial Intelligence (ICAART) e em revistas como *Expert Systems With Applications*. É um dos co-fundadores da RBCIP. Currículo lattes: http://lattes.cnpq.br/2531761427648020

5.1.2. Hércules Antonio do Prado

Atua em aplicações de sistemas baseados em conhecimento, aportando contribuições à gestão organizacional. Possui doutorado em Ciência da Computação (2001) pela UFRGS e mestrado em Engenharia de Sistemas e Computação pela COPPE/UFRJ (1989). É docente do Mestrado em Governança, Tecnologia e Inovação da Universidade Católica de Brasília, pesquisador da Rede Brasileira de Certificação, Pesquisa e Inovação (RBCIP) e integrante do seu Conselho de Administração, coordenador do Grupo de Excelência em Processo Prospectivo e Construção de Cenários do CRA-SP, Editor-Assistente da Revista Gestão do Conhecimento e Tecnologia da Informação, membro do Conselho Consultivo do Núcleo de Apoio à Pesquisa do Planejamento de Longo Prazo da FEA-USP, consultor ad-hoc da Fundação de Amparo à Pesquisa de Pernambuco (FACEPE) e da Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) e revisor técnico-científico de diversas conferências e periódicos. De 1982 a 2018, foi analista da Empresa Brasileira de Pesquisa Agropecuária, Diretor Técnico e Administrativo-Financeiro da Fundação de Apoio à Pesquisa Científica e Tecnológica (Fundação Eliseu Alves), membro do Conselho Curador da mesma fundação, membro do corpo de avaliadores de cursos de graduação do Instituto Nacional de Estudos e Pesquisas (INEP), integrante do Comitê Técnico da Embrapa Sede e Professor visitante da *School of Information Sciences* na *University of Pittsburgh*, EUA (1999). Publicou mais de 130 trabalhos científicos, entre artigos em periódicos e eventos, livros e capítulos de livros. Orientou ou co-orientou mais de 40 dissertações de mestrado e 3 trabalhos de conclusão de cursos de graduação. Curriculo lattes: http://lattes.cnpq.br/1350331210278996

5.1.3. Ítalo de Pontes Oliveira

Possui graduação no curso superior de tecnologia em Telemática pelo Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB - Campus Campina Grande), possui mestrado em Ciência da Computação pela Universidade Federal de Campina Grande (UFCG), com ênfase em Machine Learning/Deep Learning, possui diversos artigos publicados na área[4], possui publicações em revistas e conferências internacionais, como na IEEE International Symposium on Broadband Multimedia Systems and Broadcasting e IEEE Southwest Symposium on Image Analysis and Interpretation. Durante o mestrado, trabalhou em projetos de Pesquisa & Desenvolvimento em parcerias com empresas como a HP Labs e a Dell, em ambos os casos desenvolvendo soluções que usam Inteligência Artificial para resolução de problemas do mundo real e aplicados na prática no dia-a-dia dessas empresas. As soluções desenvolvidas possuem foco na área de Visão Computacional (HP) e em Processamento de Linguagem Natural (Dell). Além das experiências citadas, atuou

[4]Link para lista de publicações de Ítalo de Pontes Oliveira: https://scholar.google.com.br/citations?user=3R5wdZIAAAAJ&hl

como revisor IEEE Transactions on Affective Computing. Possui patente de uma aplicação para processamento de vídeos[5]. Participou do Hackfest promovido pelo Ministério Público da Paraíba em 2017, com o objetivo de desenvolver aplicações de análise de dados para combate à corrupção, o qual foi premiado com medalha de prata[6]. Possui cursos de formação no Coursera[7] (*TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning*) e Udemy[8] (*Spark and Python for Big Data with Pyspark*). Atualmente, atua como Cientista de Dados Líder na FabWork, projetando, desenvolvendo e gerenciando a execução de diferentes projetos usando Inteligência Artificial e Ciência de Dados aplicados nos mais diferentes contextos da economia (empresas de varejo, siderúrgicas, ministérios, entre outros). Currículo lattes: http://lattes.cnpq.br/8530154038678495.

5.1.4. Miguel Maurício Isoni

Bacharel em Administração pela FACE/FUMEC, Especialista em Contabilidade de Custos (UFPB), com curso de Marketing Internacional pela Europe University (CESED), é Mestre em Ciência da Computação pela UFPB/Campina Grande, atual UFCG, e, também, Doutor em Ciência da Informação pela Universidade Estadual de São Paulo (UNESP). Vinculado a Instituição Federal de Ensino Superior (IFES) desde 1979, exerceu vários cargos e funções, destacando ter sido: Membro do Conselho Superior da UFPB, Coordenador de Cursos de Graduação, Chefe de Departamento e Coordenador do Curso Stricto Sensu de Mestrado Profissional em Gestão nas Organizações Aprendentes (MPGOA). Exerceu a Coordenação de Equipes de Desenvolvimento de Sistemas, sendo Executivo do Centro de Processamento de Dados e dirigindo, na UFPB, a rede BITNET (uma das antecessoras da Internet). Atualmente é Professor Assistente III, exercendo seus encargos no Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba, no curso de graduação em Administração (disciplinas da Trilha de Tecnologia da Informação) e atendendo alunos do Bacharelado em Administração, Ciência da Computação e Engenharias. Sua vasta experiência em TI e Gestão de Equipes, desde 1974, o transformou em um generalista, com experiência nas seguintes empresas: TELEMIG, VALE do Rio Doce, Ferteco e PENSE Informática (automação bancária do Paraiban). Na academia escreveu capítulos de livros, artigos em periódicos nacionais e estrangeiros, participando de diversos congressos, orientando diversas pesquisas, trabalhos de conclusão de cursos e dissertações de mestrado nas seguintes áreas de interesses: (1) Melhoria Contínua de Processos; (2) Liderança, Colaboração e Empoderamento; e (3) Modelo de Negócios e Proposição de Valor Estratégico, que gerou, como projeto de pesquisa e extensão, o

---

[5]Link para Registro de Patente:
https://drive.google.com/file/d/1hgIszxpRKthN99LEtSK9V5koOlCh7h6s/view
[6]Membro do time Quebra Câmara, Quebra Senado:
https://www.gov.br/cgu/pt-br/assuntos/noticias/2017/08/hackfest-cgu-premia-aplicativos-de-combate-a-corrupcao
https://www.cnmp.mp.br/portal/todas-as-noticias/10369-primeira-etapa-do-hackfest-2017-chega-ao-fim-e-dez-equipes-se-classificam-para-as-finais
[7]Link para o certificado no Coursera:
coursera.org/verify/THPTF
[8] Link para o certificado no Udemy:
ude.my/UC-TJNZO7IG

Observatório de Startup da cidade de João Pessoa. Seu Currículo Lattes é: http://lattes.cnpq.br/8706477618884415.

5.1.5. Miguel Maurício Isoni Filho

Bacharel (2011) e Mestre com louvor (2012) em Administração com ênfase em Tecnologia da Informação e Comunicação pela Universidade Federal da Paraíba (UFPB). Cofundador & CEO da FABWORK, um centro de inovação e empreendedorismo de impacto e tecnologia. Cofundador das startups de Inteligência Artificial AxonData (2016-2018) e NEORON.io (2019-atualmente). Esta última é uma plataforma de inteligência artificial conversacional. Participante do programa global Intel AI Builders - ecossistema promovido pela Intel para acelerar a adoção de Inteligência Artificial. Ex-professor e pesquisador de Estatística Aplicada, Econometria, Métodos Quantitativos e Sistemas de Informação da Universidade Presbiteriana Mackenzie, em São Paulo. Membro-fundador e vice-líder do Global Technology, Information & Society (GTIS), grupo de pesquisa certificado pelo CNPq.

5.2. N2VEC

A N2VEC Tecnologia é uma startup de tecnologia especializada em algoritmos de Busca Inteligente e aplicações na área jurídica para Pesquisa Jurisprudencial. Desde sua fundação em 2019, a N2VEC desenvolve tecnologias inovadoras para resolução desses problemas, contando com investimento e apoio do programa PIPE da FAPESP, para empresas e projetos de alta tecnologia. Também faz parte do programa global de aceleração de startups de tecnologia no Canadá, promovido pela Dream2b e pelo Spark-Centre.

5.2.1. Fernando José Vieira da Silva

Mestre (2012) e Doutor (2020) em Ciência da Computação pelo Instituto de Computação da Unicamp, ambos na área de pesquisa de Processamento de Linguagem Natural. Possui mais de 16 anos de experiência profissional em projetos de P&D em desenvolvimento de software, sendo 8 anos em cargos de liderança. Atua como CEO e CTO na N2VEC Tecnologia, liderando pesquisas na área de Inteligência Artificial aplicada à Pesquisa Jurisprudencial em parceria com a FAPESP. Em suas experiências mais recentes, foi coordenador e professor do curso de pós-graduação/especialização em Ciência de Dados na Faculdade de Engenharia de Sorocaba e atuou como professor em cursos de graduação e em engenharia e tecnologia da mesma instituição e no curso de pós-graduação/especialização MBA em Gestão e Desenvolvimento de Sistemas em Java na Universidade de Sorocaba. Também trabalhou como Engenheiro de Processamento de Linguagem Natural na startup Biggerpan Inc., sediada em San Francisco, EUA e como Pesquisador no FIT - Flextronics Instituto de Tecnologia. Na área de Processamento de Linguagem Natural, trabalhou com mecanismos de buscas e ranqueamento utilizando aprendizado de máquina, e técnicas de reconhecimento de entidades nomeadas. Também trabalhou com pesquisas sobre resolução automática de pronomes utilizando algoritmos baseados em sumarização automática de textos e teorias do discurso e com identificação

de emoções em texto usando técnicas *cross-domain*. Currículo lattes: http://lattes.cnpq.br/6982171736310835

5.2.2. Patricia Felipe da Costa

Patricia Felipe da Costa possui graduação em Sistemas de Informação pela Universidade de Sorocaba e é pós-graduada especialista em Gestão de Projetos pelo Senac, com 8 anos de experiência profissional em tecnologia da informação, tendo atuado principalmente pela empresa Flextronics, com sede na cidade de Sorocaba. Também é estudante de Direito pela Faculdade de Direito de Sorocaba (Fadi), com mais de dois anos de experiência desenvolvendo trabalhos voluntários na área jurídica para a Prefeitura Municipal de Votorantim-SP e para o Serviço de Assistência Jurídica da Universidade de Sorocaba.

5.2.3. Pedro Henrique Correa Kim

Pedro é Engenheiro de Controle e Automação pela Unesp e Especialista em Ciência de Dados pela Faculdade de Engenharia de Sorocaba. Atualmente também é Mestrando em Ciências da Computação pela Ufscar, com ênfase em Inteligência Artificial. Também possui 2 anos de experiência profissional como Engenheiro de Machine Learning.

5.3. RBCIP

A Rede Brasileira de Certificação, Pesquisa e Inovação - RBCIP é uma associação civil com personalidade jurídica de direito privado, sem fins econômicos, estatutariamente e legalmente (lei 13.243/16) enquadrada como instituição científica, tecnológica e de inovação (ICT), sua finalidade é fomentar e promover o ensino, a pesquisa científica, o desenvolvimento tecnológico e o desenvolvimento institucional. Somos reconhecidos pela Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) como Organização da Sociedade Civil capaz de usufruir dos benefícios previstos na Lei nº 8.010, de 29 de março de 1990, alterada pela Lei nº 10.964, de 28 de outubro de 2004.

A Rede Brasileira de Certificação, Pesquisa e Inovação conta com diversos parceiros estratégicos nacionais e internacionais, tais como a Universidade Federal de Mato Grosso do Sul, Universidade Federal de Ouro Preto, Laboratório de Inteligência Pública da Universidade de Brasília - UnB. O quadro social da RBCIP é formado por membros associados efetivos e associados fundadores os quais podem ser visto por meio do link https://www.rbcip.org/pesquisadores

Os associados efetivos que demonstram interesse em participar do projeto e que podem contribuir para alcançar o objetivo proposto são os seguintes:

5.3.1. Arthur Mesquita Camargo

Doutorando em Contabilidade pelo PPGCont-UNB. Mestre em Administração Pública UnB (2014). Especialista em Contabilidade Pública (UNISUL/2014) e Finanças Públicas (ESAF/2011). Tem mais de 10 anos de experiência no Setor Público, sendo responsável pelo desenvolvimento de soluções tecnológicas por Microsoft Azure voltadas

para conselhos de fiscalização de atividade profissional, em específico, no Controle Contábil, Orçamentário e Despesas, Portal da Transparência, Relatório de Gestão do TCU, Controle de Prestação de Contas. Atualmente, é pesquisador sênior na RBCIP e Diretor Administrativo e Financeiro, exercendo a Coordenação dos projetos Laboratório de Apoio à Inovação da Educação Básica do Brasil (LabInova) e Laboratório de Criatividade da Educação Básica (LabCrie). Tem pesquisas na área pública na análise de conteúdo. Curriculo lattes: http://lattes.cnpq.br/1195882649429046

5.3.2. Marcelo Estrela Fiche

Doutor em Economia aplicada pela UNB, Mestre em Economia pela Universidade Federal de Santa Catarina e graduado na Escola de Formação de Oficiais da Marinha Mercante (1992) com especialização em máquinas. Especialista em Direito Tributário e Finanças Públicas pelo Instituto de Direito Público. Auditor Federal de finanças e Controle - Secretaria do Tesouro Nacional / Ministério da Fazenda (1995 - atual). Ocupou cargos gerenciais tais como: Gerente de arrecadação da ANVISA, Assessor econômico do Ministro do Conselho de Desenvolvimento Econômico e Social - CDES da Presidência da Republica, Coordenador-Geral de execução financeira do FNDE/MEC, Coordenador-Geral de Arrecadação do Salário-Educação, Coordenador-Geral de Política Fiscal e Chefe de gabinete da Secretaria de Política Econômica do Ministério da Fazenda, Assessor Especial e Chefe de Gabinete do Ministro da Fazenda. Pesquisador Associado no Centro de Estudos Avançados de Governo e Administração Pública da Universidade de Brasília (CEAG/FACE). Como docente foi professor e coordenador dos cursos de Ciências Econômicas e coordenador e professor do Programa de Pós-Graduação Stricto Sensu em Governança, Tecnologia e Inovação. Diretor Presidente da Rede Brasileira de Certificação, Pesquisa e Inovação. Coordenou projeto ligados a redes complexas e aprendizado de máquina: o papel da interconectividade - RCAM. É integrante da Plataforma de Ciência de Dados aplicada às Políticas Públicas, do Laboratório de ciência de dados aplicada a Economia e Governança. Tem publicações na área de machine learning. Curriculo lattes: http://lattes.cnpq.br/4282659017553803

5.3.3. Lilian Campos Soares

Mestre em Gestão do Conhecimento e Tecnologia da Informação (MGCTI) pela Universidade Católica de Brasília (2018). Especialista em Ciência de Dados e Big Data pela Pontifícia Universidade Católica de Minas Gerais (2019/2020) e em Redes de Computadores pela Universidade Católica de Brasília (2000). MBA em Gestão Tecnológica e de Negócios pela Universidade Federal do Rio de Janeiro (2002). Bacharel em Informática pela Universidade de Fortaleza (1997). Pós-Graduanda em Inovação e Gestão Ágil de Projetos pelo IGTI (2021). Experiência na Ciência da Computação, com ênfase em Infraestrutura e Governança de TI, Planejamento Estratégico de TI, Planejamento Diretor de TI, Gestão de Projetos de TI, Gestão de Sistemas de Informação. Planejamento de Transporte e Logística. Desenvolvimento, implantação e operação de Observatórios. Plataformas de Analytics e BI, Ciência de Dados e serviços de TIC em nuvem. Possui sólidos conhecimentos da Gestão da Logística Pública (Licitação, Contratos, Materiais, Suprimentos, Convênios para Concedentes), Planejamento Governamental, Orçamento Público e Gestão Orçamentária e Financeira. Atualmente é Coordenadora do Observatório Nacional de Transporte e Logística da EMPRESA DE PLANEJAMENTO E LOGÍSTICA S.A.

Tem pesquisas na área de blockchain e redes convergentes. Curriculo lattes: http://lattes.cnpq.br/2617785793194225

5.3.4. Thiago Christiano Silva

Em 2012, obteve o título de Doutor em Ciências Matemáticas e de Computação pelo Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). Em 2009, completou a graduação em Engenharia de Computação pela Universidade de São Paulo, formando-se em primeiro lugar da turma. Sua tese de doutorado recebeu três prêmios acadêmicos, entre eles o "Prêmio Capes de Tese 2013", outorgado pela Capes. Publicou o livro "Machine Learning in Complex Networks" pela Springer em 2016 (já conta com tradução em chinês). Possui bolsa de Produtividade em Pesquisa Nível 2 na área de Administração, Contabilidade e Economia do CNPq (2020 - 2022) e também coordena projeto da Universal do CNPq (2019 - 2021) na área de Ciência da Computação. O artigo "Stochastic Competitive Learning in Complex Networks" publicado no IEEE Transactions on Neural Networks and Learning Systems foi artigo destaque pela IEEE Computational Intelligence Magazine (vol. 7, no. 3, 2012). É professor doutor da Universidade Católica de Brasília (UCB) e integra dois programas de pós-graduação: (i) o programa de Doutorado/Mestrado em Economia (linha de Finanças) e (ii) e o programa de Mestrado Profissional em Governança, Tecnologia e Inovação (linha de Ciência de Dados e Aprendizado de Máquina). Atua também como pesquisador e chefe de divisão da Consultoria de Pesquisa em Estabilidade Financeira no Departamento de Estudos e Pesquisas (Depep) do Banco Central do Brasil (BCB). Possui interesse em tópicos relacionados a Ciência de Computação, Finanças e Economia. Trabalha com: aprendizado de máquina, redes complexas, estabilidade financeira, risco sistêmico, econometria e banking. Coordenador das seguintes pesquisas: Efeitos da Interconectividade na Economia: uma abordagem microeconômica e com redes complexas e Finanças e Crescimento: uma abordagem com redes complexas e aprendizado de máquina. Possui mais extensa publicação na área de machine learning. Curriculo lattes: http://lattes.cnpq.br/6238208958412798

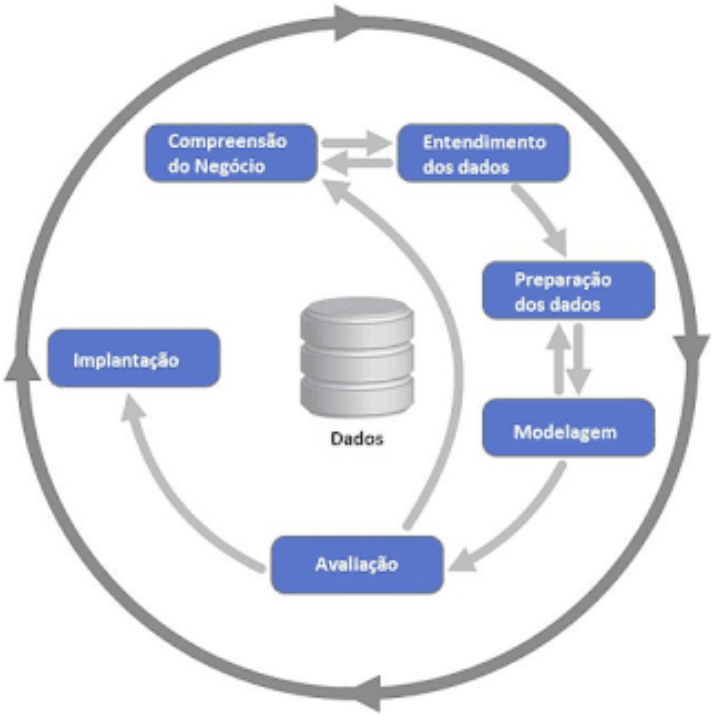
5.3.5. Benjamin Miranda Tabak

Professor da Escola de Políticas Públicas e Governo da Fundação Getulio Vargas (FGV EPPG). Tem experiência em Economia e Direito, com ênfase em Regulação Financeira, Análise Econômica do Direito e Análise Econômica do Direito Comportamental. Pesquisa sobre Economia Bancária, Finanças e Direito, com artigos científicos e livros publicados nessas áreas e em assuntos correlatos. Tem trabalhos publicados na área de Machine Learning e Deep Learning. Editor associado de revistas especializadas nacionais e estrangeiras. Bolsista de Produtividade em Pesquisa do CNPq - Nível 1B. Doutorado e Mestrado em Economia. Professor da Fundação Getúlio Vargas e Consultor legislativo no Senado Federal. Possui publicação na área de machine learning. Curriculo lattes: http://lattes.cnpq.br/7238063563586831

6. Metodologia

6.1. CRISP-DM

A metodologia adotada para o projeto é a *Cross Industry Standard Process for Data Mining* (CRISP-DM). O modelo de ciclo de vida do CRISP-DM consiste em seis fases com setas indicando as dependências mais importantes e frequentes entre as fases:



Passos da CRISP-DM.

Fonte: LEMOS, Jorge Luiz Cavalcante.

Figura - Representação do ciclo do CRISP-DM.

1. **Compreensão do Negócio:** nesta fase, deve ser explorado o que a organização espera ganhar com o projeto de mineração de dados. Recomenda-se envolver nas discussões pessoas diretamente afetadas pelo projeto ou detentoras de conhecimento.
2. **Compreensão dos Dados:** esta fase corresponde ao estágio de familiarização com os dados do problema e identificação da qualidade dos mesmos, obtenção das primeiras percepções e formulação de prognoses sobre o que os dados podem mostrar.
3. **Preparação dos Dados:** realiza-se tarefas como seleção e integração de tabelas, amostragem, criação de novos atributos, limpeza dos dados brutos iniciais,

construção de gráficos, elaboração do dicionário de variáveis com seus respectivos tipos e particionamento do arquivo em dados de treinamento e dados de teste.

4. **Modelagem:** nesta fase é construído um modelo sobre os dados, conforme o tipo de tarefa a ser realizada (agrupamento, classificação, associação, previsão, etc) por um algoritmo de aprendizagem de máquina.
5. **Avaliação:** neste momento, é importante avaliar e rever os passos executados para a obtenção do modelo que permitirá o alcance dos objetivos do projeto. Resultados insatisfatórios acarretarão o retorno à fase inicial do processo para sua reestruturação.
6. **Implementação:** esta é a fase em que o conhecimento adquirido é organizado, apresentado e colocado em uso. Corresponde à aplicação dos novos insights para fazer melhorias na organização. Isso pode significar a criação de novos processos ou a integração formal do modelo criado a algum processo existente.

O trabalho será desenvolvido remotamente pelo consórcio de empresas, com interações virtuais com a equipe do Tribunal de Contas da União quando a atividade envolver *brainstorming*, *mentoring*, especificações, avaliações, comunicações de resultados e operação em produção. Todas as atividades realizadas serão documentadas e estão contidas no relatório final a ser entregue.

7. Bibliografia

DAI, Andrew M.; OLAH, Christopher; LE, Quoc V. Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998, 2015.

McCormick, C. 2016. Word2vec Tutorial – The Skip-Gram Model. Disponível em http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model . Acesso em 08 de Fevereiro de 2021.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” ICLR (2013).

PERONE, Christian S. (2013, December, 09). Machine Learning :: Cosine Similarity for Vector Space Models (Part III). Retrieved from http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Shperber, Gidi (2017, July, 26). A gentle introduction to Doc2Vec. Retrieved from https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e