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

Flavio Augusto Weber

Técnicas de Processamento de Linguagem Natural Aplicadas à Gestão de Serviços de TI

Curitiba
2020

Flavio Augusto Weber

Técnicas de Processamento de Linguagem Natural Aplicadas à Gestão de Serviços de TI

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

Orientador: Walmes Zeviani

Curitiba

2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Técnicas de Processamento de Linguagem Natural Aplicadas à Gestão de Serviços de TI

Flavio Augusto Weber[1]
Maurício Ferste[2]
Walmes Zeviani[3]

Resumo

Em muitas empresas, que prestam serviços de Tecnologia da Informação (TI), usuários e clientes solicitam as mais diversas demandas envolvendo em especial serviços de infraestrutura. Nesse processo, um dos principais problemas é a categorização inicial indevida da demanda, gerando "desvios" relacionados à falta de informação por parte do usuário, cliente ou, ainda, atendente, no momento do registro da mesma. Assim sendo, a categorização deve ser corrigida e encaminhada para o grupo de atendimento correto. Neste artigo, serão apresentadas técnicas de Processamento de Linguagem Natural (PLN), bem como algoritmos de Aprendizado de Máquina Supervisionados baseados em múltiplas classes, os quais, juntos, oferecem soluções de Inteligência Artificial para a classificação automática de textos, tornando as atividades de categorização de demandas de TI mais eficientes, de forma a propor uma solução para o problema de classificação acima citado.
Palavras-chave: Aprendizagem de Máquina, Estatística, Inteligência Artificial, Mineração de Dados, Mineração de Textos, Processamento de Linguagem Natural (PLN).

Abstract

In many companies that works with Information Technology (IT), users and customers request the most diverse demands involving, in particular, infrastructure services. One of the main problems is the improper initial categorization of demand, generating mistakes related to the lack of information on the part of the user, customer or, even, attendant, at the time of registering it. Therefore, the categorization must be corrected and forwarded to the correct service group. In this article, Natural Language Processing (NLP) techniques will be presented, as well as Supervised Machine Learning algorithms based on multiple classes, which together offer Artificial Intelligence solutions for automatic text classification, making categorization activities more efficient IT demands, in order to propose a solution to the classification problem above.
Keywrods: *Machine Learning, Statistics, Artificial Intelligence, Python Programming Language, Data Mining, Text Mining, Natural Language Processing (NLP).*

1 Introdução

Muitas empresas que prestam serviços de TI, entre elas o **SERPRO – Serviço Federal de Processamento de Dados**, uma empresa do Ministério da Economia, a qual possui Regionais e Escritórios presentes em todo o território nacional, oferecendo soluções digitais para o Estado, Sociedade e cidadãos brasileiros, atendem usuários e clientes, os quais solicitam as mais diversas demandas envolvendo, por exemplo, serviços de infraestrutura como rede local, segurança, certificação digital, correio eletrônico, entre outros. Geralmente, essas demandas são solicitadas através de um sistema de acionamentos no qual o próprio usuário ou cliente descreve sua necessidade em campos de um formulário online. Nesse momento, a demanda deve ser associada a uma determinada categoria, que, por sua vez deverá, então, estar relacionada ao seu respectivo grupo de atendimento. O registro inicial da demanda também pode ser realizado por um atendente, através de uma central de atendimento ou, ainda, através de um processo automatizado de atendimento, como um *chatbot* [4].

Nesse processo, um dos principais problemas é a categorização inicial da demanda, pois geralmente ocorrem "desvios" durante o direcionamento desta ao seu grupo de atendimento. Esses desvios estão relacionados à falta de informação por parte do usuário ou cliente ou, ainda, por parte do atendente, quando do preenchimento da demanda em questão. Assim sendo, a categorização deve ser corrigida e encaminhada para o grupo de atendimento correto. Outro problema que ocorre é a demora no atendimento da demanda, já que o processo se torna muitas vezes moroso devido à correção do encaminhamento ao respectivo grupo de atendimento. De modo a resolver essas questões, uma solução seria categorizar de

[1]Aluno do programa de Especialização em Data Science & Big Data, flavioaw@gmail.com.
[2]Analista de Sistemas do SERPRO, mauricio.ferste@serpro.gov.br.
[3]Professor do Departamento de Estatística - DEST/UFPR.

[4]Programa de computador que tenta simular um ser humano na conversação com as pessoas.

forma automática e o mais assertivo possível as demandas aos seus respectivos serviços, de modo a evitar recategorizações e reduzir os desvios de encaminhamentos das mesmas, com o intuito de diminuir sensivelmente o tempo de atendimento ao usuário ou cliente. Para tal, podemos recorrer à utilização de *Data Mining*[5] para extrair e preparar os dados, juntamente com técnicas de *Natural Language Processing*[6] (Processamento de Linguagem Natural) e de algoritmos de *Machine Learning*[7] (Aprendizado de Máquina) supervisionados baseados em múltiplas classes com fins a construir uma solução de Inteligência Artificial (IA).

2 Fundamentação Teórica

Ultimamente, temos visto um número crescente de aplicações que armazenam dados não estruturados devido ao avanço acelerado das tecnologias de informação. Dados esses, que na maioria das vezes, incluem informações valiosas, como por exemplo: tendências, anomalias e padrões de comportamento que podem ser utilizados para auxiliar nas tomadas de decisões ((BRITO, 2017)). Em razão disso, variadas técnicas foram desenvolvidas com o propósito de recuperar informações importantes contidas em bases de dados, originando a área denominada *Text Mining* (Mineração de Textos), a qual é derivada das técnicas de Data Mining (Mineração de Dados).

Segundo Turban et al. ((2019)), **mineração de dados** é o processo de identificação de padrões válidos, novos, potencialmente úteis e definitivamente compreensíveis em dados armazenados em bases de dados estruturados, onde os dados encontram-se organizados em registros estruturados por variáveis categóricas, ordinais ou contínuas. Ressalta que é um processo que emprega técnicas estatísticas, matemáticas e de inteligência artificial para extrair e identificar informações úteis e conhecimentos (ou padrões) a partir de vastos conjuntos de dados. Turban et al. (2) Turban et al. ((2019)) destaca ainda que mineração de dados é o mesmo que **mineração de textos**, visto que ambos processos têm o mesmo propósito e utilizam os mesmos processos, porém, no caso da mineração de texto, a entrada ao processo se dá através de uma coleção de arquivos de dados não estruturados. As tarefas de mineração de dados (ou mineração de textos), geralmente, são classificados em três categorias principais: previsão, associação e agrupamento. Essa categorização dependerá da forma como os padrões dos dados serão extraídos e determinará o tipo de algoritmo de aprendizado de máquina a ser utilizado: supervisionado ou não supervisionado. Nos algoritmos de aprendizado supervisionado os dados de treinamento incluem os atributos descritivos (variáveis independentes ou variáveis decisórias) e o atributo de classe (variável de saída ou de resultado), já no caso do aprendizado não supervisionado, os dados de treinamento incluem apenas os atributos descritivos ((Turban et al., 2019)).

Dentre os algoritmos da categoria "previsão", encontram-se os de classificação, regressão e série temporal. Em se tratando de dados não numéricos, utilizamos os algoritmos de mineração de dados do tipo classificação, já que os algoritmos de regressão tratam dados numéricos. Os algoritmos de classificação analisam os dados armazenados numa base de dados e geram um modelo capaz de prever comportamentos futuros, consistindo em generalizações baseadas nos registros de um conjunto de dados de treinamento que ajudam a distinguir classes pré-definidas ((Turban et al., 2019)).

O processo de mineração de textos é composto por um conjunto de etapas conforme proposto por ((Aranha and Vellasco, 2007)):

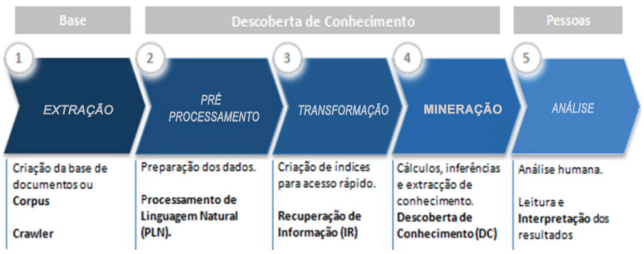

Figura 1: Diagrama ilustrativo da metodologia de mineração de textos

Essa metodologia apresenta cinco etapas, sendo a primeira a coleta de dados, a segunda o pré-processamento desses dados, a terceira etapa consiste na criação de índices que possibilitam melhor desempenho na recuperação dos dados, a quarta se concentra na aquisição de conhecimento e finalmente a quinta fase dedicada à interpretação dos resultados obtidos.

Na etapa de extração (coleta de dados), cria-se uma base de dados textual. Essa base de dados constituirá o "corpus" onde serão aplicadas as técnicas de mineração de textos. Um corpus (plural corpora) é um conjunto vasto e estruturado de textos, geralmente armazenados e processados eletronicamente, preparado com o propósito de conduzir descoberta de conhecimento ((Turban et al., 2019)).

A fase de pré-processamento é iniciada após o término da coleta de dados. Nela é despendido, provavelmente, a maior parte do tempo consumido no processo. Consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair ((BRITO, 2017)). Essa fase compõe o "*pipeline*" dos processos de PLN (Processamento de Linguagem Natural).

O processamento de linguagem natural (PLN) é uma técnica importante utilizada na mineração de textos. Conforme Miranda, utilizando-se dos conhecimentos da área

[5]Processo de explorar grandes quantidades de dados à procura de padrões consistentes.

[6]Subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais.

[7]Subárea da inteligência que estuda meios para que máquinas posam fazer tarefas que seriam executadas por pessoas.

de linguística, o PLN permite aproveitar ao máximo o conteúdo do texto, extraindo entidades, seus relacionamentos, detectando sinônimos, corrigindo palavras escritas de forma errada e ainda não ter ambiguidade. Para Turban et al. ((2019)), o PLN é um componente da mineração de texto, além de ser um subdomínio da inteligência artificial e da linguística computacional. Turban et al. ((2019)) complementa que o PLN estuda o problema da "compreensão" da linguagem humana natural, com a visão de converter retratos da linguagem humana (como documentos textuais) em representações mais formais (na forma de dados numéricos e simbólicos) que sejam mais fáceis de manipular em programas de computador. Na visão de Siegel ((2017)), o PLN é o campo de pesquisa que desenvolve a tecnologia para trabalhar com a linguagem humana, destacando que:

"Se os dados são a água da Terra, o dado textual é a parte conhecida como o oceano. Considerados como 80% de todos os dados, eles são tudo que nós, a humanidade, sabemos e que nos preocupamos em escrever. É muito poderoso – e rico em conteúdo porque foi criado com a intenção de propagar não apenas fatos e números, mas conhecimento humano. Mas texto, a maior oportunidade dos dados, apresenta o maior dos desafios".

BRITO ((2017)) salienta a existência de seis técnicas de PLN conforme o nível linguístico processado: fonológico (lida com a pronúncia), morfológico (trata as palavras isoladamente), lexical (trabalha com o significado das palavras), sintático (refere-se à estrutura das frases), semântico (interpreta os significados das frases) e pragmático. Dentre estas, sobressai-se a morfológica. A abordagem de dados seguindo essas técnicas também podem ser consideradas como uma análise semântica. Uma outra abordagem utilizada é a análise estatística, na qual a importância de um termo é dada pelo número de vezes que este aparece no texto ((Morais and Ambrosio, 2007)). Ambas abordagens podem ser utilizadas separadamente ou em conjunto. Para o presente artigo, serão abordadas as técnicas relacionadas ao nível morfológico juntamente com a análise estatística.

Bases textuais contém uma grande quantidade de termos e atributos para sua representação, resultando em uma denotação esparsa, com muitos atributos nulos. Assim sendo, complementa BRITO ((2017)), as técnicas de pré-processamento são importantes na medida em que resolvem os problemas decorrentes dos dados textuais, identificando melhor os atributos que representam o conhecimento, reduzindo drasticamente a quantidade destes, sem perder as características principais da base de dados. Durante a etapa inicial, são realizadas tarefas básicas de limpeza de texto, como conversão de texto em minúsculo, remoção de números, remoção de pontuação e remoção de espaços em branco, de modo a não alterar de modo geral o significado do texto, visto que para o processamento do mesmo para fins de aprendizado de máquina o que importa são as palavras mais significantes presentes no texto.

Em seguida, realiza-se a tokenização do texto, a qual consiste na conversão de uma cadeia de caracteres de entrada em uma cadeia de palavras ou *tokens*, sendo que os delimitadores utilizados para a tokenização geralmente são o espaço em branco entre os termos, quebras de linhas, tabulações e alguns caracteres especiais.
A terceira técnica é a remoção das palavras irrelevantes – *stopwords*, pois são termos que não agregam nenhum valor, além de aumentarem o tempo de processamento, como conectores textuais, conjunções, preposições, interjeições e artigos.
No processo de normalização de palavras, há duas técnicas: *stemming* e lematização. No caso do *stemming*, ocorre a redução de palavras à raiz ou à sua forma básica. Já na lematização, acontece a redução das formas flexionais a uma forma de base comum, ou seja, ao contrário do *stemming*, a lematização não simplesmente corta as inflexões. Ela usa bases de conhecimento lexicais para obter as formas básicas corretas de palavras (raiz da palavra – lema). Exemplificando:

Tabela 1: Lematização x *Stemming*.

Palavra	Lematização	*Stemming*
Estudar	Estudar	Estud
Estudou	Estudar	Estud
Estudando	Estudar	Estud

Definido o tipo de abordagem de dado a ser utilizado no processo de mineração de textos, segue-se para a etapa de preparação de dados. A preparação dos dados faz parte da fase de pré-processamento dos dados, onde é realizado o processo de descoberta de conhecimento em textos, no qual estão envolvidos a seleção dos dados que constituem a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo destes textos. Como resultados dessa etapa temos: provimento de uma **redução dimensional**, identificação de similaridades em função da morfologia ou do significado dos termos nos textos ((Morais and Ambrosio, 2007)). Assim sendo, após as tarefas de pré-processamento, pode-se utilizar a técnica de "*bag-of-words*", que significa "saco de palavras", sendo uma forma de extrair recursos do texto para ser utilizado pelos algoritmos de aprendizado de máquina. Nesse modelo, textos, como uma frase, um parágrafo ou um documento completo são representados como uma coleção de palavras, ignorando a gramática ou a ordem em que as palavras aparecem ((Turban et al., 2019)). O propósito é descobrir a frequência de cada *token*, também conhecido como Frequência de Termos, chegando ao conceito de TF-IDF. TF-IDF significa frequência de termo de frequência inversa de termo, sendo que seu peso é uma medida usada para avaliar a importância de uma palavra para um documento em um corpus e sua importância aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, porém é compensada pela frequência da palavra no corpus. Temos então o TF (Frequência do Termo) que é uma pontuação da frequência da palavra do documento atual e o IDF (Frequência Inversa de Documentos) que é a pontuação de quão rara a palavra é entre documentos.

Ainda, durante a preparação dos dados, são realizadas tarefas de normalização que removem o que for desnecessário para a compreensão do texto bem como prepara-o para que o mesmo seja analisado e classificado da melhor forma possível, e, segundo Miranda, nessa etapa de pré-processamento, é realizado um conjunto de transformações sobre a coleção de textos de modo que passem a ser estruturados em uma representação atributo-valor, o qual possa ser manipulada pelos métodos de extração de conhecimento, sendo que a obtenção de tal representação pode ser feita através da realização de algumas tarefas como identificação dos atributos, atribuição de pesos e redução da representação.

Segundo BRITO ((2017)), na etapa de mineração são aplicadas técnicas direcionadas ao aprendizado de máquina a fim de obter novos conhecimentos. Dessa forma, os algoritmos a serem utilizados dependerão da necessidade de informação do usuário. Os algoritmos de aprendizado de máquina possuem, dentro do contexto de classificação automática de textos, o objetivo de aprender, generalizar ou ainda extrair padrões ou características das classes dos textos de uma coleção, afirma ((Junior and Rossi, 2017)). BRITO ((2017)) ressalta que os algoritmos desenvolvidos sobre a aprendizagem de máquina se baseiam na estatística e na probabilidade para aprender padrões complexos a partir de algum corpus. Para o objetivo deste artigo os modelos de aprendizado de máquina serão aqueles orientados à supervisão, pois trata-se de dados que serão classificados de acordo com classes pré-estabelecidas. Assim sendo, os algoritmos de aprendizagem supervisionados que serão utilizados neste artigo são: *Naive Bayes*, *Linear SVC*, *Logistic Regression* e *Random Forest*.

Naive Bayes é um algoritmo bseado na teoria das probabilidades que, segundo Harrison ((2020)), pressupõe uma independência entre os atributos dos dados, ou seja, os atributos irão influenciar a classe de forma independente. Conforme descreve Amaral ((2016)), este algoritmo classificador constrói uma tabela mostrando o quanto cada categoria de cada atributo contribui para cada classe. Em seguida, ao submeter uma nova instância para o classificador, o mesmo verifica os pesos na tabela, soma-os e analisa qual classe tem peso maior, o qual se tornará o vencedor nesta classificação.

O algoritmo ***Linear SVC*** (Máquinas de Vetores de Suporte Lineares) é derivado do modelo SVM – *Support Vector Machines* (Máquina de Vetores de Suporte), podendo ser utilizado para classificações lineares envolvendo dados complexos, como textos. Para Harrison ((2020)), o SVM é um algoritmo que tenta fazer a adequação de uma linha (ou plano ou hiperplano) entre as diferentes classes de modo a maximizar a distância da linha até os pontos das classes, procurando encontrar uma separação robusta entre as classes. Nesse contexto, os vetores de suporte são os pontos da fronteira do hiperplano divisor.

Logistic Regression (Regressão Logística) é outro algoritmo de classificação que estima probabilidades de uma instância pertencer a uma determinada classe usando uma função logística. A função logística é uma função sigmoide (em formato de S) que mostra um número entre 0 e 1, assim sendo se a probabilidade estimada for maior que 50% (cinquenta por cento), então o modelo prevê que a instância pertence a essa classe (chamada classe positiva e rotulada como "1"), caso contrário, prevê que não (classe negativa e rotulada "0") (11).

Random Forest (Floresta Aleatória) é um algoritmo baseado em outro algoritmo supervisionado de classificação chamado Árvores de Decisão (*Decision Tree*). Uma árvore de decisão usa uma estrutura de árvore para representar um número de possíveis caminhos de decisão ((Grus, 2019)). Nesse caso, *Random Forest* é um conjunto de árvores de decisão e um resultado para cada caminho. Segundo Harrison ((2020)), a ideia das florestas aleatórias é criar uma floresta de árvores de decisão treinadas em diferentes colunas dos dados de treinamento. Se cada árvore tiver uma chance melhor que 50% (cinquenta por cento) de fazer uma classificação correta, a predição será incorporada. O algoritmo *Random Forest* é baseado em previsões de um conjunto de previsores, resultando muitas vezes em melhores previsões do que com o melhor previsor individual. Um conjunto de previsores é chamado de *ensemble*; assim, esta técnica é chamada *Ensemble Learning*, e um algoritmo de *Ensemble Learning* é chamado de *Ensemble Method* ((Geron, 2019)).

Na fase de análise é realizada a interpretação dos resultados, sendo avaliada a eficiência do processo de mineração de textos após a obtenção dos dados gerados ao final da aplicação dos algoritmos de aprendizado de máquina. Em outras palavras, é o momento de avaliar se o objetivo foi cumprido da melhor forma possível, descobrindo conhecimento novo e inovador a partir de pilhas de documentos não-estruturados ((Miranda)).

Assim sendo, nessa etapa é feita a avaliação dos classificadores, considerando, a princípio, os seguintes resultados possíveis para os documentos classificados ((Kultzak, 2016)):

Tabela 2: Classificações utilizadas na matriz de confusão.

VP – Verdadeiros Positivos	O documento foi classificado corretamente em relação a uma categoria
FP - Falsos Positivos	O documento está relacionado a uma categoria de forma incorreta
FN - Falsos Negativos	O documento não está relacionado a uma categoria mas deveria estar
VN - Verdadeiros Negativos	O documento não deveria estar relacionado a determinada categoria e de fato não está.

VP e VN são as classificações corretas. Numa matriz de confusão, essas classificações aparecem na diagonal principal.

2.1 Métricas

Esses quatro possíveis resultados possibilitam estabelecer métricas para avaliação de classificação: acurácia, *recall*, precisão, F1-*score*.
A **acurácia** (*accuracy*) é a porcentagem de classificações corretas - a razão entre o somatório das previsões corretas (positivos com verdadeiros negativos) sobre o somatório das previsões, ou seja, diz o quanto o modelo acertou das previsões possíveis:

$$\text{Acurácia: } A = \frac{TP + TN}{TP + FP + TN + FN}.$$

Recall (revocação), também conhecido como sensibilidade (*sensitivity*), é a porcentagem de valores positivos classificados corretamente. Em outras palavras, quão bom é o modelo para prever positivos, sendo positivo a classe que se quer prever. É determinado como a razão entre verdadeiros positivos sobre a soma de verdadeiros positivos com negativos falsos:

$$\text{Recall: } R = \frac{TP}{TP + FN}.$$

A **precisão** (*precision*) é a porcentagem de predições positivas que estavam corretas, ou seja, o quão bem o modelo se comportou:

$$\text{Precisão: } P = \frac{TP}{TP + FP}.$$

F1-*score* é a média harmônica de recall e precisão. A pontuação F1 pode ser uma medida melhor quando se precisa buscar um equilíbrio entre precisão e *recall* e houver uma distribuição de classe desigual (grande número de verdadeiros negativos - TN):

$$\text{F1: } F = 2 \cdot \frac{P \cdot R}{P + R}.$$

A estimação da acurácia de um modelo de classificação induzido por um algoritmo de aprendizado de máquina é importante por dois motivos: primeiro, pode ser usado para estimar futura precisão preditiva, sugerindo o nível de confiança depositado nas previsões geradas pelo classificado; segundo, pode ser usado para a seleção de um classificador para um determinado conjunto. Uma metodologia para estimativa muito utilizada é a **validação cruzada *k-fold***. Essa metodologia, também chamada de estimativa por rotação, procura minimizar a tendenciosidade associada ao fracionamento aleatório das amostras de dados de treinamento e de reserva ao se comparar a precisão preditiva de dois ou mais métodos. Nessa metodologia, o conjunto completo de dados é dividido de forma aleatória em k subconjuntos mutuamente excludentes de tamanho praticamente igual e o modelo de classificação é treinado e testado *k vezes*. A cada vez, o modelo é treinado em todos, exceto um dos subconjuntos e então testado no único subconjunto restante ((Turban et al., 2019)).
Ressalta-se que o processo de mineração de textos é cíclico, ou seja, ao término de cada fase os resultados precisam ser analisados e caso seja necessário, realizam-se modificações no processo, iniciando um novo ciclo.

3 Métodos

3.1 Ambiente

Para realizar as tarefas de processamento computacional necessárias para a mineração de textos, foi utilizado a linguagem *Python*, a qual está sendo utilizada atualmente em larga escala para atividades de programação, especialmente àquelas ligadas ao aprendizado de máquina e inteligência artificial. Foram utilizadas, bibliotecas específicas dessa linguagem, como o NLTK – *Natural Language Toolkit*, contendo programas para processamento de linguagem natural e o *SCIKIT-LEARN*, que contém programas de aprendizado de máquina. Esses softwares foram executados no *Jupyter Notebook*, que é um ambiente de desenvolvimento web.

3.2 O conjunto de dados

Para dar início ao processo de mineração de textos e, por conseguinte, às demais fases do mesmo, foi construída uma base de dados através da coleta de dados. Os dados coletados foram obtidos da base de dados do sistema de gerenciamento de serviços automatizado da empresa (SERPRO). Esses dados foram transformados numa planilha eletrônica, constituída de 1992 (mil novecentos e noventa e três) linhas (registros), cada uma contendo 2 (dois) atributos: "Demandas" e "Categoria". A coluna Demandas contém a descrição da própria demanda informada pelo usuário, e, a coluna Categoria, a respectiva categorização. Assim sendo, considera-se o conjunto de dados (*dataset*) já pré-classificado nas diversas categorias de demandas.

Tabela 3: Exemplo de conjunto de dados (*dataset*) utilizado.

	CATEGORIA	DEMANDA
0	Certificação	'Prezados, necessitamos de...
1	Segurança	(29005/2016) Solicito o desbl...
2	LAN	(31946/2019) - Solicito um rel...
...	...	...
1989	LAN	Verificar no log do relay mail-...
1990	LAN	Verificar relay smtp mail-apl...
1991	LAN	Visando atender à integração...

3.3 Análise exploratória

Como já mencionado, na fase de pré-processamento é realizado um conjunto de transformações de modo a melhorar a qualidade dos dados disponíveis e organizá-los, visando à etapa de mineração, onde serão submetidos à aplicação de algoritmos de aprendizagem de máquina. Abaixo, temos a limpeza básica do texto, removendo pontuações, números, caracteres especiais, além do emprego das técnicas de tokenização, *stemming* e lematização. A seguir o exemplo da aplicação das técnicas numa determinada demanda de usuário:

A Josefina, responsável pela operacionalização das emissões de certificados na STN, nos solicitou documentação sobre procedimento a seguir para emitir certificados digitais para a STN. Esta documentação deve também incluir os acordos operacionais realizados entre SERPRO e STN, a saber: Ao vir ao SERPRO, a instalação do certificado no token é feita pela equipe de certificação digital do mesmo, ou seja, os usuários devem voltar para a STN com o certificado já instalado no token. Quando tratar-se de usuários VIP, cuja assinatura é coletada na STN, a instalação do certificado no token ficará sob responsabilidade da COSIS. Foi solicitado também incluir no documento os requisitos para utilização do sistema web.

Figura 2: Conteúdo de uma demanda antes do pré-processamento do texto.

josefina responsavel operacionalizacao emissoes certificados stn solicitou documentacao sobre procedimento seguir emitir certificados digitais stn documentacao deve tambem incluir acordos operacionais realizados stn saber vir instalacao certificado token feita equipe certificacao digital usuarios devem voltar stn certificado ja instalado token tratar usuarios vip cuja assinatura coletada stn instalacao certificado token ficara sob responsabilidade cosis solicitado tambem incluir documento requisitos utilizacao sistema web

Figura 3: Conteúdo de uma demanda após do pré-processamento do texto

Observando a figura 3, nota-se a redução de palavras da demanda em questão. De 110 (cento e dez) palavras, obtém-se 62 (sessenta e dois) tokens, ou seja, uma diminuição de 56% (cinquenta e seis por cento), após a aplicação das técnicas de pré-processamento, as quais facilitam o processamento dos textos e a otimização dos modelos de aprendizagem de máquina.
Analisando o conjunto de dados, composto pelas demandas e por suas respectivas categorias, verifica-se a quantidade de demandas por categoria:

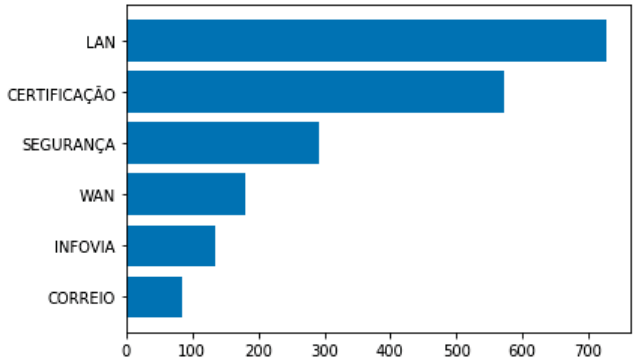

Figura 4: Gráfico de barras ilustrando as categorias e respectivas quantidades

Percebe-se, no gráfico da figura 4, a predominância das categorias “Certificação” e “LAN”.

Abaixo a nuvem de palavras, que mostra as palavras mais significativas (maior ocorrência) no corpus analisado:

Figura 5: Nuvem de palavras representativas do “corpus” analisado

4 Resultados e Discussões

O conjunto de dados foi treinado e testado 10 (dez vezes) através da técnica de validação cruzada *k-fold* (k=10) para os algoritmos de aprendizado de máquina (*Linear SVC*, *Logistic Regression*, *Naive Bayes* e *Random Forest*), obtendo-se as seguintes acurácias:

Tabela 4: Classificações utilizadas na matriz de confusão.

Modelo	Acurácia
Linear SVC	0.845915
Logistic Regression	0.802711
Naive Bayes	0.769601
Random Forest	0.598379

O desempenho de cada modelo também pode ser observado através do gráfico:

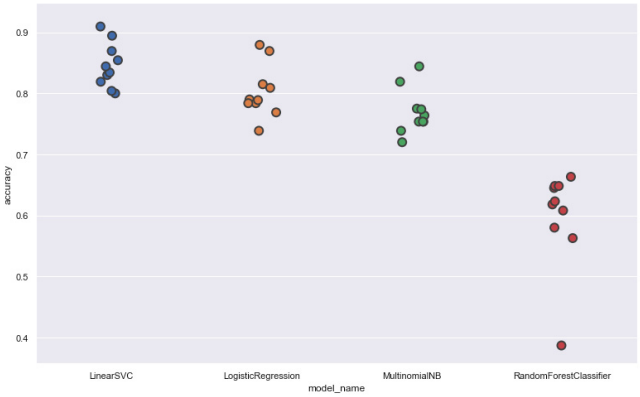

Figura 6: Gráfico de dispersão demonstrando o desempenho dos modelos de classificação.

Cada modelo de classificação apresenta 10 (dez) pontos no gráfico de dispersão representando cada treinamento

e teste realizado para o mesmo.
Abaixo, a matriz confusão do modelo *Linear SVC*, que obteve a melhor pontuação de acurácia:

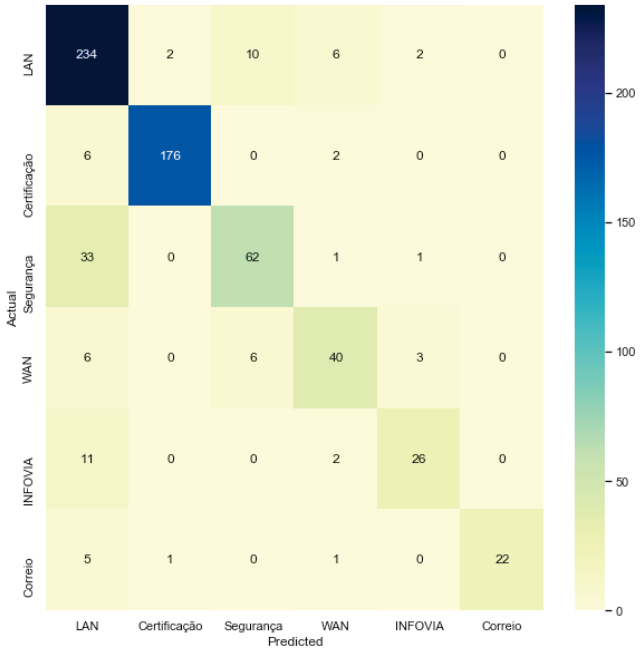

Figura 7: Matriz de confusão resultante do modelo *Linear SVC*

A matriz confusão demonstra, para o modelo *Linear SVC*, as diferenças entre os rótulos previstos e reais.
A seguir, as métricas para o modelo *Linear SVC*:

Tabela 5: Métricas obtidas da aplicação do classificador *Linear SVC*.

Categoria	Precisão	*Recall*	F1-*Score*
Certificação	0.98	0.98	0.98
INFOVIA	0.87	0.62	0.72
Correio	0.86	0.68	0.76
LAN	0.84	0.88	0.86
WAN	0.83	0.87	0.82
Segurança	0.78	0.87	0.82
Acurácia	-	-	0.87
Média Total	0.86	0.80	0.82
Média Ponderada	0.87	0.87	0.87

5 Análise

Após a execução dos algoritmos de aprendizado de máquina, verifica-se que o modelo *Linear SVC* teve um desempenho muito bom, 84% (oitenta e quatro por cento) seguido do modelo *Logistic Regression* que também teve uma boa resposta, obtendo uma acurácia de 80% (oitenta por cento).

A classificação correta das categorias (precisão) obteve uma média total de 86% (oitenta e seis por cento), enquanto a sua média ponderada foi de 87% (oitenta e sete por cento). O recall, que demonstra o modelo em termos de previsão de valores positivos registrou média ponderada de 87% (oitenta e sete por cento). Finalmente, o F1-score, que representa a média entre o recall e a precisão, equilibrando a distribuição de classes, alcançou média total de 82% (oitenta e dois por cento) e ponderada de 87% (oitenta e sete por cento).

6 Conclusão

As métricas de classificação de desempenho obtidas pelo modelo *Linear SVC*, para o conjunto de dados utilizado (tabela 3), juntamente com as técnicas de processamento de linguagem natural utilizadas, como limpeza do texto, remoção de stopwords, tokenização, stemming e lematização, demonstram que o modelo se comportou muito bem, mesmo para uma distribuição não equilibrada de dados, conforme verificado na figura 4, sendo que nesse caso, muitas vezes possa ser interessante contar com um classificador que forneça alta precisão de predição para classes majoritárias, ao passo que mantém uma precisão razoável para as classes minoritárias. Dessa forma, fica constatado que é possível solucionar problemas de categorizações de textos, como o apresentado neste artigo, otimizando a classificação das categorias das demandas de forma inteligente e automatizada, utilizando técnicas de processamento de linguagem natural e algoritmos de aprendizado de máquina, fornecendo um modelo eficiente de classificação multiclasse.

Referências

F. Amaral. *Aprenda Mineração de Dados: Teoria e Prática*. Alta Books, 2016.

C. N. Aranha and M. Vellasco. *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. PUC-RIO, 2007.

E. M. N. D. BRITO. *Mineracão de textos: detecção automática de sentimentos em co-mentários nas mídias sociais. Projetos e Dissertacões em Sistemas de Informação e Gestão do Conhecimento*. 2017.

A. Geron. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn Tensorflow*. Alta Books, 2019.

J. Grus. *Data Science do Zero: Primeiras Regras com o Python*. Alta Books, 2019.

M. Harrison. *Machine Learning – Guia de Referência Rápida*. Novatec, 2020.

D. Junior and R. Rossi. *Classificação automática de textos utilizando aprendizado supervisionado baseado em uma única classe. Trabalho de conclusão de curso de Sistemas de Informação*. UFMS-CPTL, 2017.

A. F. Kultzak. *Categorização de textos utilizando algoritmos de aprendizagem de máquina com WEKA*. 2016.

A. R. O. Miranda. *Descoberta de Conhecimento em texto aplicada a um sistema de atendimento aos usuários de um plano de assistência à saúde*. URL http://www.coc.ufrj.br/pt/dissertacoes-de-mestrado/109-msc-pt-2009/1588-aline-regina-de-oliveira-miranda.

E. Morais and A. Ambrosio. *Mineraçao de textos. Relatório Técnico–Instituto de Informática*. UFG, 2007.

E. Siegel. *Análise Preditiva. O poder de prever quem vai clicar, comprar, mentir ou morrer*. Alta Books, 2017.

E. Turban, R. Sharda, and D. Delen. *Business Intelligence e Análise de Dados para Gestão do Negócio*. Bookman, 2019.