

# Categorização de Perguntas para SAC Utilizando Mineração de Dados

*Data Mining Applied to Questions Categorization in Customer Service Systems*

**Matheus Pinto**<sup>1</sup>

 [orcid.org/0000-0003-1143-3525](https://orcid.org/0000-0003-1143-3525)

**Richard Rocha**<sup>1</sup>

 [orcid.org/0000-0002-8825-0455](https://orcid.org/0000-0002-8825-0455)

**Murilo Stodolni**<sup>1</sup>

 [orcid.org/0000-0002-0875-2907](https://orcid.org/0000-0002-0875-2907)

**Nilton Silva**<sup>1</sup>

 [orcid.org/0000-0002-9361-0267](https://orcid.org/0000-0002-9361-0267)

<sup>1</sup>Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: [mpap@ecomp.poli.br](mailto:mpap@ecomp.poli.br)

**DOI: 10.25286/rep.v7i3.2449**

Esta obra apresenta Licença Creative Commons Atribuição-Não Comercial 4.0 Internacional.

Como citar este artigo pela NBR 6023/2018: Matheus Pinto; Murilo Stodolni; Nilton Silva; Richard Rocha. Categorização de Perguntas para SAC Utilizando Mineração de Dados. Revista de Engenharia e Pesquisa Aplicada, Recife, v. 7, n. 3, p. 10-18.

## RESUMO

A situação de receber perguntas através de e-mails ou portais desenvolve uma problemática latente que é a falta de tempo para desenvolvimento de respostas em Ouvidorias, sendo este um dos grandes desafios de empresas que insere essa forma de comunicação com seus clientes. O presente artigo tem como objetivo utilizar técnicas de Mineração de Dados para classificar as perguntas recebidas.

No estudo de caso deste artigo, serão utilizadas as técnicas de Mineração de Dados para auxiliar a tomada de decisões de processos administrativos e também a manipulação de um sistema de Perguntas e Respostas frequentes da Secretaria da Controladoria Geral do Estado. Ao final desse artigo, ficará evidente as análises das diversas técnicas utilizadas na construção do sistema, como também os resultados por elas obtidos. Dessa forma, a proposta de uma mudança para um ambiente automatizado poderá auxiliar o fluxo de informações e de execução de tarefas dentro da Controladoria.

**PALAVRAS-CHAVE:** Mineração de Dados; Extração de Características; Perguntas Frequentes;

## ABSTRACT

The situation of receiving questions through e-mails or portals develops a latent problem that is the lack of time to develop answers in Ombudsman services, which is one of the great challenges of companies that insert this form of communication with their customers. This article aims to use Data Mining techniques to classify incoming questions.

In the case study of this article, text data mining techniques will be used to aid decision making in administrative processes and also the manipulation of a Frequently Asked Questions and Answers system of the Secretary of the Comptroller General of the State. At the end of this article, the analysis of the various techniques used in the construction of the system will be evident, as well as the results they sought. In this way, a proposal for a change to an automated environment can help the flow of information and the execution of tasks within the Controllorship.

**KEY-WORDS:** Data Mining; Features Extraction; Frequently Asked Questions;

## 1 INTRODUÇÃO

Com a crescente necessidade de digitalização [1] de processos e serviços das empresas, também surgiu a necessidade de que processos correlacionados com o suporte de suas soluções se adaptem ao cenário digital. No início da pandemia de COVID 19, o processo foi acelerado. Cerca de 40% dos diretores alavancaram a digitalização tendo como objetivo expandir o crescimento operacional [2]. Tal setor recebeu pontos positivos ao ser capaz de se utilizar de um canal de comunicação mais simplificado e de amplo acesso, sendo possível fornecer um atendimento personalizado para cada cliente.

É comum que as empresas e organizações tenham um serviço de atendimento ao consumidor (SAC). Em 2020, em pesquisa realizada pela Zendesk [3] constatou que 75% dos clientes se sentem mais confiantes e satisfeitos ao adotar serviços que possuam uma boa experiência ao longo da utilização; melhorando o atendimento, aumentando a produtividade e reduzindo o custo operacional.

### 1.1 DESCRIÇÃO DO PROBLEMA

As empresas continuam atentas às novas necessidades dos seus clientes. No entanto, a resolução da maioria das dúvidas inseridas na plataforma é feita de forma manual e pode demorar semanas até ser respondida. Essa problemática gera um atraso na entrega de respostas, sendo latente a necessidade de uma alternativa que auxilie o departamento responsável. Trata-se de um processo laborioso que envolve a identificação de similaridades com as dúvidas já solucionadas e classificação da natureza das novas dúvidas.

Segundo o Instituto Ibero-Brasileiro de Relacionamento com o Cliente [4], demonstra que a ausência de automatização e rapidez durante os atendimentos têm aumentado. Em 2014, cerca de 39% dos consumidores estavam insatisfeitos com a resolução dos seus problemas/dúvidas por meio dos canais de atendimento. Em 2016, o número aumentou para 71%. Tal problemática é oriunda da má estruturação dos dados e do trabalho manual na extração de características em um "mar de palavras".

### 1.2 OBJETIVOS

Nesse direcionamento, o trabalho tem como objetivo prover a classificação de perguntas através de assuntos, por meio de implementação de técnicas de mineração textual de dados, evoluindo assim, a extração de características dos textos de orientação.

### 1.3 JUSTIFICATIVA

Os sistemas manuais de atendimento ao consumidor tornam as soluções para as dúvidas um trabalho demorado e dispendioso. Essa demora impacta diretamente os usuários do sistema em suas tomadas de decisão, diminuindo assim a eficiência dos agentes [3] e uma possível perda de receita para a empresa.

A partir de um novo classificador que contribua para a rápida determinação do assunto tratado em cada pergunta inserida permite que a realocação de recursos materiais e profissionais seja mais dinâmica e eficiente.

### 1.4 ESCOPO NEGATIVO

Fica evidente, portanto, que não se trata do escopo deste presente trabalho a elaboração de um novo sistema, nem da resolução automática das perguntas, mas sim a melhoria das funcionalidades preexistentes no âmbito atual, em especial no que diz respeito à extração de características e tratamento de anomalias através de um novo motor classificador dos assuntos das perguntas.

## 2 FUNDAMENTAÇÃO TEÓRICA

### 2.1 OUVIDORIA

Durante o decorrer dos anos, muitas sociedades viviam sob um sistema onde não existia uma situação de pertencimento social quando se tratava de limitações da administração pública. Para sanar esse impasse, muitos países criaram uma condição de participação social sobre essas limitações, dessa forma a palavra *Ombudsman* (no português, Ouvidoria) começou a percorrer sobre os ambientes de situação pública com a função de investigar desavenças, reclamações, sugestões, possibilitando assim construir soluções para as demandas sociais. De acordo com o jornal científico *European*

*Scientific*, especificamente o artigo *Ombudsman - Historical Views* [5], a origem da Ouvidoria Pública vem da Suécia, e tem como significado *representante do povo*, sendo o funcionário cujo recebia o título, a pessoa que deveria ser responsável para receber as discussões da população, objetivando agir em discussões juntamente ao Parlamento.

Ouvidoria é a função do controle interno, que tem por finalidade fomentar o controle e a participação de indivíduos externo a organização, por meio do recebimento, registro e tratamento de denúncias e manifestações sobre os serviços prestados à sociedade e a adequada aplicação dos serviços, visando a melhoria da sua qualidade, eficiência e resolubilidade, tempestividade e equidade.

Esses sistemas de ouvidoria, são por muitas vezes, utilizados de forma manual, para atender as demandas solicitadas pela população. Uma alternativa adotada por órgãos públicos é utilizar um canal que proponha a inclusão de FAQs, de forma a otimizar as respostas aos cidadãos, mesmo que por muitas vezes sejam ineficientes.

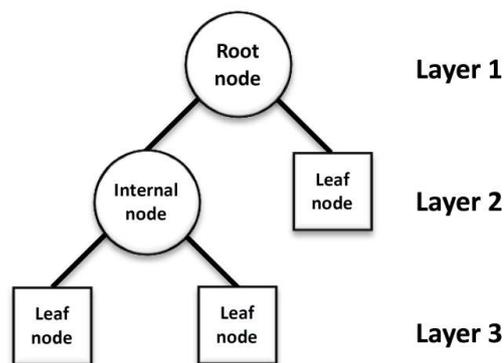
## 2.2 ÁRVORES DE DECISÃO

A ideia de classificação está presente desde a antiguidade, onde Aristóteles propunha um sistema de classificação de forma a organizar e sistematizar o conhecimento apresentando uma primeira sistemática de classificação dos animais [8]. Sendo assim, a computação se utiliza também dessas ideias de forma a reunir dados com características semelhantes e separação dos não afins.

O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema (Breiman *et al.*, 1984). Para o processo de classificação, existem vários algoritmos dentro do ramo de inteligência artificial que atendem a esse propósito. Dentro do ponto de vista de classificação textual as árvores de decisão se apresentam como uma boa aplicação.

A árvore de decisão é um modelo representado graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido (Han & Kamber, 2001; Monard & Baranauskas, 2002a; Witten & Frank, 2005). O nó raiz é o primeiro nó da árvore, no topo da estrutura. Os nós internos, incluindo o nó raiz, são nós de decisão. Cada um contém um teste sobre uma variável independente e os resultados desse teste formam os ramos da árvore.

Figura 1 - Exemplo de uma árvore de decisão.



Fonte: o autor.

### 2.2.1 Extra Trees Classifier

Além da precisão, a capacidade de generalização é de fundamental importância para o estudo de caso em questão. Tal classificador gera múltiplas árvores de decisão aleatórias e então, combinando os resultados de cada árvore, encontra a resposta final [8]. Sua principal diferença é que o processo é extremamente aleatório, contribuindo para a geração de padrões generalizados.

De maneira análoga ao Random Forest, o algoritmo realiza uma seleção aleatória das variáveis candidatas ao nó inicial, os dados presentes em cada uma dessas variáveis serão segregados (divisão de dados) também de forma aleatória.

## 2.3 TRABALHOS RELACIONADOS

Tendo em vista o presente cenário de um ambiente de perguntas e respostas, muito já se tem pesquisado em termos de inteligência artificial e mineração de dados para extração de informações e melhoria das respostas para diminuição da atividade manual, dentre esses trabalhos podemos destacar BADA e MENEZES (2012), que trata do esclarecimento de dúvidas por meio de assistentes pessoais, SOUZA (2009) que propõe um sistema de resposta automática para perguntas fechadas através de cálculos de similaridade e proximidade, e também PRESTES (2011) avaliação de diferentes algoritmos utilizados na seleção da resposta.

### 2.3.1 Análise e Classificação de Textos

Neste trabalho, verifica-se a análise de texto, que é voltada para a análise lexical, sendo implementada uma abordagem estatística, que visa extrair as palavras mais relevantes de um determinado texto. Portanto, o artigo procurou demonstrar que usando soluções de PLN, os agentes podem melhorar significativamente sua tarefa de identificar textos relevantes para a pesquisa usando dados espaciais (web) como referência de pesquisa. O processamento do conteúdo textual proposto compreende três etapas: manipulação, análise e classificação da linguagem natural [6].

O trabalho dá ênfase para a necessidade utilização de técnica de tokenização, em que o texto será transformado em tokens, remoção de *stopwords* e inclusive a normalização dos tokens para caracteres minúsculos. O processo é justificado visto que os tokens representam a base de significação das palavras envolvidas, sendo chamadas de morfemas lexicais.

### 2.3.2 Classificação Textual utilizando Deep Learning

Mineração de texto é um processo de descoberta de conhecimento que usa técnicas para analisar e extrair dados de texto, frases ou palavras. É o processo de extrair padrões ou conhecimentos interessantes e não triviais de documentos em textos não estruturados (Sebastiani 2002). Esta descoberta de conhecimento envolve diversas aplicações como análise de texto, extração de informação, síntese, classificação, clustering, linguística computacional, entre outras.

Nesse sentido, torna-se evidente que "Compreender" um texto significa perceber o contexto, realizar análises sintáticas, semânticas, lexicais e morfológicas, gerar resumos, extrair informações, interpretar significados, analisar sentimentos, até aprender conceitos de textos processados [7].

### 2.3.3 Aplicação de Técnicas de Mineração de Texto para Classificação de Sentenças Judiciais Brasileiras

O trabalho trata da aplicação da automação especialmente no que se refere à classificação das penas judiciais em diferentes categorias, especializado em "ler" os textos e classificá-los com base em textos passados cujo conteúdo foi inspecionado, total ou parcialmente, por humanos.

Baseado em técnicas de mineração de texto e processamento de linguagem natural (PNL), o sistema proposto neste estudo constitui uma estrutura de análise de dados capaz de ser metodologicamente refinada e estendida a aplicações mais gerais, fornecendo índices e dados. Os indicadores podem ser usados posteriormente para, por exemplo, medir o desempenho jurídico no Brasil, o que por sua vez pode ajudar a moldar políticas públicas para melhorar o judiciário brasileiro.

## 3 MATERIAIS E MÉTODOS

### 3.1 STAKEHOLDERS ENVOLVIDOS

No que se refere aos stakeholders envolvidos, seus nomes e posições são descritos no Quadro 1.

**Quadro 1** - Descrição dos stakeholders envolvidos.

STAKEHOLDER	POSIÇÃO
Renato Cirne	Consultor de Privacidade de Dados na SER.Educacional e ex-Gestor governamental de Controle Interno da SCGE
Helder Silva	Colaborador Tático e Operacional do Núcleo de Inovação e ICT
João Amaral	Diretor de Tecnologia de Informação do Controle Interno

**Fonte:** Dados fornecidos pela SCGE.

### 3.2 DESCRIÇÃO DA BASE DE DADOS

No Quadro 2 estão listadas as principais colunas da base de dados, sendo estas as utilizadas para o desenvolvimento da categorização proposta.

**Quadro 2** - Descrição acerca da composição da base de dados utilizada na construção do estudo de caso.

CAMPO	DESCRIÇÃO	TIPO
codigo_pergunta	Cada pergunta do demandante é registrada no sistema com uma chave única no sistema	int

## Categorização de Perguntas para SAC Utilizando Mineração de Dados

resposta_publica	Resposta do especialista no sistema	string
unidade_gestora	Nome da unidade gestora demandante da pergunta	string
assunto	Classificação do assunto realizada pelo especialista	string
resposta	Situação final da demanda entre "respondida" e "não respondida"	string
id_pergunta	Ordem numérica das perguntas	int
cod_ug	Código da unidade gestora	int

**Fonte:** Dados fornecidos pela SCGE.

### 3.3 ANÁLISE DESCRITIVA DOS DADOS

A partir dos dados obtidos pela base da Lei de Acesso à Informação, foram computadas a quantidade de registros feitos por ano solicitando atendimento pela controladoria.

#### 3.3.1 Distribuição de Frequência

**Quadro 3 -** Distribuição de frequência dos assuntos.

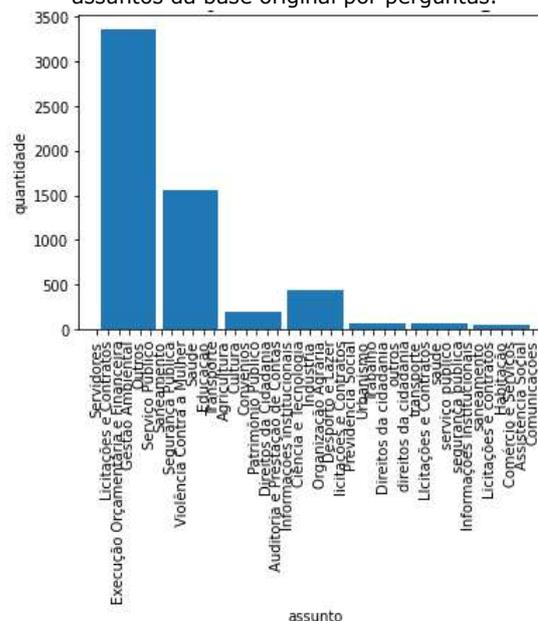
ASSUNTO	<i>fi</i>	<i>fri</i>	<i>Fac</i>	<i>FRi</i>
Servidores	924	24,31%	924	24,31%
Licitações e contratos	440	11,56%	1364	35,89%
Execução orçamentária e financeira	405	10,66%	1769	46,55%
Gestão ambiental	244	6,42%	2013	52,97%
Saneamento	155	4,08%	2168	57,05%
Violência contra a mulher	152	4%	2320	61,05%
Informações institucionais	242	6,37%	2562	67,42%
Segurança pública	449	11,81%	3011	79,24%
Outros	789	20,76%	3800	100%

**Fonte:** Dados fornecidos pela SCGE

**Legenda:** *fi* é referente à Frequência Absoluta, *fri* é referente à Frequência Relativa, *Fac* é referente à Frequência absoluta acumulada e *FRi* é referente à frequência relativa acumulada.

Para melhor visualização, o histograma a seguir mostra a quantidade de pedidos feitos relacionado a cada assunto.

**Figura 2 -** Histograma que relaciona a quantidade de assuntos da base original por perguntas.



**Fonte:** Dados fornecidos pela SCGE

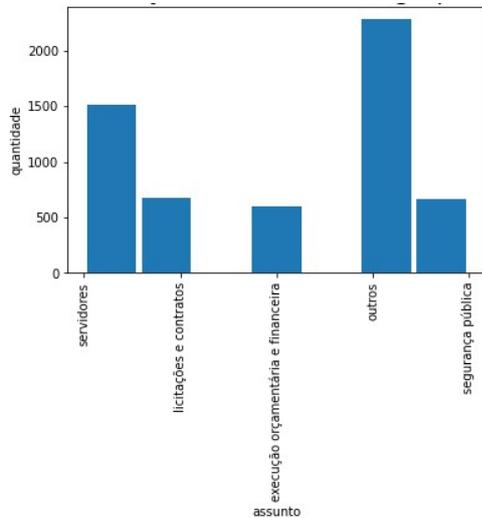
### 3.4 PRÉ-PROCESSAMENTO DOS DADOS

Durante a etapa de pré-processamento foram realizadas limpezas a fim de tornar a etapa de classificação mais assertiva - retirando toda a acentuação, caracteres especiais. Logo em seguida foi necessário retirar ruídos dos assuntos que poderiam dificultar a classificação, dessa maneira, os assuntos que tinham recorrência menor que 6% foram agrupados como assunto "outros" (Figura 3).

Posteriormente uma lematização (tratamento que mantém apenas os prefixos das palavras) foi realizada (Figura 4).

Por fim, foi realizada a remoção das palavras menos relevantes para os assuntos, deixando as 50 (cinquenta) palavras mais relevantes por assunto.

**Figura 3** - Distribuição dos assuntos após a remoção dos assuntos com baixa ocorrência.



Fonte: Os autores

**Figura 4** - Nuvem de palavras lematizadas.



Fonte: Os autores

**Figura 5** - Base de dados após processamento inicial.

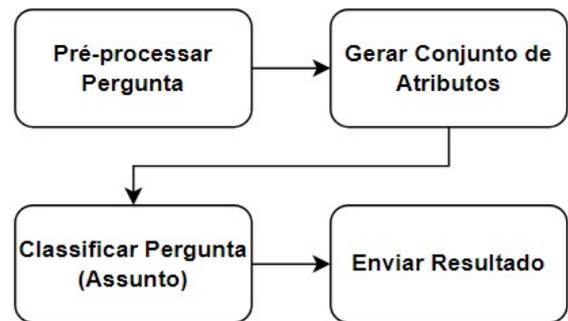
PEDIDO	ASSUNTO
solic ouvid segunt informacoes relaca contrat dat inici termin disciplin	Servidores
relaca professor efet contrat are tecnolog informaca	Servidores
relaca tod professor regim contrat temporari lingu portugues lot recif de	Servidores
solic relaca professor contrat municipi surubimc pernambuc disciplin ei	Servidores
solic relaca professor educaca fisic contrat limoeiro pernambuc	Servidores
solic secretari fazend copi contrat financ firm entr est pernambuc calx	Licitações e Contratos
solic secretari fazend copi extrat cont unic banc agenci refer mé novei	Execução Orçamentária e Financeira

Fonte: Os autores

### 3.5 METODOLOGIA EXPERIMENTAL

Para o desenvolvimento completo da metodologia experimental é necessário que todo o pipeline da Figura 6 seja implementado. Porém, o objetivo deste trabalho é realizar apenas a "Classificação Assunto".

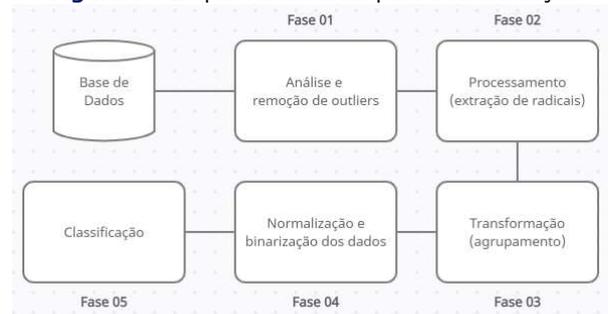
**Figura 6** - Sintetização das etapas



Fonte: Os autores

Destarte, a solução proposta abaixo segue as seguintes etapas listadas na Figura 7. Sendo assim, se faz necessária a descrição técnica de cada uma delas a seguir.

**Figura 7** - Etapas necessárias para a classificação



Fonte: Os autores

#### 3.5.1 Análise e Remoção de Outliers

Tal etapa é de fundamental importância visto que será determinante para a qualidade final dos dados que serão utilizados no processamento - impactando no modelo de previsão gerado a partir dos dados. A etapa se dividiu em dois momentos: um manual e outra que foi automatizada utilizando as operações implementadas no *DataFrame* da biblioteca *Pandas*; uma estrutura de dados muito similar a uma convencional tabela utilizada em processadores de texto do mercado.

O processo considerou realizar uma limpeza inicial dos dados, removendo registros com valores nulos (como, por exemplo, na coluna "ASSUNTO") e linhas duplicadas. Também foi levado em conta a existência de dados ruidosos, onde foram aplicadas expressões regulares a fim de atenuar as suas ocorrências ao longo da base de dados.

### 3.5.2 Extração de Radicais

Inicialmente foi proposto utilizar o processo de “stemização”, trata-se de um conjunto de algoritmos com regras para a extração de radicais sem o conhecimento do contexto envolvido. A técnica se destacou inicialmente por ser uma operação de processamento de texto rápida e fácil de se utilizar para a recuperação de informação e, inclusive, classificação. No entanto, para o problema em questão, ficou evidente que o significado das palavras em um dado contexto deve ser levado em conta; isto é, as raízes morfológicas das palavras.

Dessa forma, foi preciso aplicar o uso da “lematização” - técnica na qual determina a forma canônica da palavra como lema (forma raiz) com base num dicionário ou lista de vocabulários informados. Foi utilizada a classe *WordNetLemmatizer*, do pacote *nltk*. Para o dicionário em questão, foi utilizado a lista de vocabulário contida em *pt\_core\_news\_sm*, inclusa na biblioteca *spaCy*.

Além disso, o pacote *nltk* também foi útil na remoção das *stopwords* de cada registro do *data frame* em questão. O conjunto de palavras removido nesta fase refere-se às palavras consideradas irrelevantes na língua portuguesa no contexto da classificação.

### 3.5.3 Transformação dos Dados

Tal processo consiste na aplicação de uma série de operações, utilizando as principais técnicas de transformação com o intuito de adequar os dados, seus formatos e seus valores; tendo em vista garantir uma melhor compreensão e qualidade final para a etapa de classificação. Os processos utilizaram as operações padrões do *data frame*.

#### 3.5.3.1 Renomeação de Assuntos com Frequência Inferior ao Parâmetro Definido

Com base nos resultados obtidos na análise descritiva dos dados, ficou notória a baixa incidência de um conjunto de assuntos na base de dados em relação a outros que se demonstraram predominantes (Figura 2).

No entanto, por se tratar de uma classificação puramente textual, baseada nos radicais extraídos, se verificou que a elevada quantidade de informações envolvidas nesses assuntos continua a ser relevante para a geração das *features* no modelo classificador. Sendo assim, foi desenvolvido

um processo de transformação que visa renomear os assuntos que estejam com frequência inferior a um limiar definido (atualmente definido como 6%).

#### 3.5.3.2 Criação de *Data Frames* por Assunto

O processo visou manipular os dados criando *dataframes* auxiliares com registros de agrupados por assunto. Tal manipulação foi de fundamental importância para o entendimento e extração dos principais radicais utilizados por assunto - o qual formarão os nós do modelo classificador.

#### 3.5.3.3 Geração de *Word Clouds*

O processo de geração de *word clouds* será de vital importância para a extração dos conjuntos de radicais mais relevantes por assunto. É possível definir uma frequência base para extração e também uma quantidade máxima de *features* por assunto.

Os radicais extraídos são fundidos, e é criado um *data frame* com as *features* (radicais) e o devido rótulo (assunto). Que será utilizado para na etapa de binarização.

#### 3.5.3.4 Binarização das *Features*

A binarização consiste numa técnica que converte atributos contínuos ou discretos em binários. Por se tratar de uma classificação textual, a binarização verificará a ocorrência de cada uma das *features* extraídas com cada registro do *data frame* processado.

#### 3.5.3.5 Normalização

Como pode ser visto na Figura 3, certos assuntos continuam em escalas diferentes mesmo após uma redução inicial. Dessa forma, como recomenda a literatura, foi necessário realizar uma transformação a fim de evitar que um assunto predomine sobre o outro. Contudo, existem situações em que tal variação deve ser preservada visto que é relevante para o modelo a ser gerado. Com o *dataset* envolvido no presente artigo, a normalização dos dados trouxe uma variação percentual na métrica de precisão em torno de 2%, para mais ou para menos.

### 3.5.4 Classificação

Com o *data frame* com as *features* binarizadas, em seguida é realizada a classificação utilizando o algoritmo *ExtraTreeClassifier* (*sklearn.ensemble*). A escolha do algoritmo tem em vista garantir a capacidade de generalização, evitando o *overfitting* com a base de dados. Os dados são divididos na proporção de 75% para treino e 25% para testes.

#### 4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Após a aplicação de todas as técnicas demonstradas no decorrer do trabalho e tendo em vista atingir um maior acerto dos verdadeiros positivos, foi adotada como principal métrica a *accuracy*. Foi possível, dessa forma atingir 91% de acurácia para a classificação dos cinco assuntos propostos (Figura 8).

**Figura 8** - Resultados obtidos utilizando o algoritmo Extra Tree Classifier.

Accuracy estimated: 91.28%

	precision	recall	f1-score
execução orçamentária e financeira	0.91	0.85	0.88
licitações e contratos	0.93	0.94	0.93
outros	0.90	0.93	0.92
segurança pública	0.89	0.86	0.87
servidores	0.94	0.91	0.93
accuracy			0.91
macro avg	0.91	0.90	0.91
weighted avg	0.91	0.91	0.91

Fonte: Os autores

Na Figura 8 são apresentados os resultados da classificação alcançada para cada um dos cinco assuntos disponíveis, nessa imagem detalha a precisão alcançada além de outras duas métricas dando maior confiabilidade do modelo de árvore de decisão alcançado, como pode ser visto na Figura 9.

A métrica de *recall* é bastante útil para minimizar os casos de falsos negativos e com ela é possível calcular a métrica *f1-score* do modelo que combina precisão e *recall* de modo a trazer um número único que indique a qualidade geral do modelo e que consegue trabalhar bem até mesmo com conjuntos de dados que possuem classes desproporcionais.

**Figura 9** - Matriz de confusão do modelo criado.



Fonte: Os autores

#### 5 CONCLUSÕES E TRABALHOS FUTUROS

##### 5.1 CONCLUSÕES

Com base nos resultados obtidos, ao se utilizar da aplicação das técnicas comumente utilizadas na mineração de textos, foi possível atingir o objetivo do trabalho ao desenvolver um novo motor classificador com a alta precisão de 91% gerando uma maior confiabilidade provendo uma rápida determinação do assunto tratado em cada pergunta.

Fica evidente, portanto, que é possível se adicionar uma ferramenta que faça a primeira triagem e direcione o pedido para o setor correto, melhorando assim os processos e diminuindo o tempo de resposta para o cliente.

##### 5.2 TRABALHOS FUTUROS

Como trabalhos futuros, é notória a necessidade de uma base com uma maior quantidade de assuntos, para que não seja necessário agrupar os assuntos com poucas ocorrências em "outros" aumentando a quantidade de rótulos classificáveis.

Tendo em vista a completude de toda a aplicação utilizada pela controladoria, seria ideal a realização de esforços em pesquisa e desenvolvimento dos seguintes motores: identificador de similaridades entre os pedidos, gerador de tópicos, incluindo a construção de uma plataforma web que integre as soluções desenvolvidas.

#### REFERÊNCIAS

- [1] PERINI, Mário A.; LIBERATO, Yara G.; SARAIVA, Maria F.; FULGÊNCIO, Lúcia. **Sobre a classificação de palavras**. Disponível em:

<https://www.scielo.br/j/delta/a/F6VF5tdyFN7zGKZLWfJ7PXP>. Acesso em: 01 dez 2021.

- [2] PWC. **Global Digital Trust Insights**. Disponível em: [2021 Global Digital Trust Insights: PwC](#). Acesso em: 25 out 2021.
- [3] ZENDESK. **Customer experience trends**. Disponível em: <https://www.zendesk.com.br/customer-experience-trends/>. Acesso em: 21 out 2021.
- [4] INSTITUTO IBERO-BRASILEIRO DE RELACIONAMENTO COM O CLIENTE. **Pesquisa revela aumento de insatisfação com 'call centers'**. Disponível em: <http://g1.globo.com/bom-dia-brasil/noticia/2016/02/pesquisa-revela-aumento-de-insatisfacao-com-call-centers.html>. Acesso em: 20 out 2021.
- [5] Hajdari, Erika; University, Tirana, Albania. **OMBUDSMAN- HISTORICAL VIEWS**. Disponível em: <https://eujournal.org/index.php/esj/article/view/2860>. Acesso em: 03 nov 2021.
- [6] BAGATINI, Daniela; FARINON, Joao Luis. **Análise e classificação de conteúdo textual**. Disponível em: [https://online.unisc.br/acadnet/anais/index.php/salao\\_ensino\\_extensao/article/view/14110](https://online.unisc.br/acadnet/anais/index.php/salao_ensino_extensao/article/view/14110). Acesso em: 15 nov 2021.
- [7] SANTOS, Keila. **Categorização de textos por aprendizagem de máquina**. Disponível em: <https://keilabcs.github.io/publications/dissertacao.pdf>. Acesso em: 16 nov 2021.
- [8] BROWNLEE, Jason. **How to develop an Extra Trees Ensemble with Python**. Disponível em: <https://machinelearningmastery.com/extra-trees-ensemble-with-python>. Acesso em: 16 nov 2021.