

**FACULDADE DE TECNOLOGIA ALCIDES MAYA - AMTEC  
CURSO TECNOLÓGICO EM SISTEMAS PARA INTERNET**

**DIEGO PORTELA MARINHO**

**DETECÇÃO DE SPAM ATRAVÉS DE APRENDIZADO DE MÁQUINA**

**Porto Alegre  
2019**

**DIEGO PORTELA MARINHO**

**DETECÇÃO DE SPAM ATRAVÉS DE APRENDIZADO DE MÁQUINA**

Projeto de Pesquisa apresentado como requisito parcial para obtenção do título de Tecnólogo em Sistemas para Internet, pelo Curso de Sistemas para Internet da Faculdade de Tecnologia Alcides Maya - AMTEC

Orientador: Prof.Me. Maicon dos Santos

Porto Alegre

2019

## LISTA DE SIGLAS

---

AM	Aprendizado de Máquina
API	Interface de Programação de Aplicação

---

## LISTA DE FIGURAS

1 – Fluxo de dados do método de aprendizado.....	20
2 – Código em Python do método de aprendizado.....	20
3 – Fluxo de dados do método de detecção.....	22
4 – Código em Python do método de detecção.....	22
5 – Código em Python do método de tratamento de dados.....	23
6 – Código em Python utilizado para gerar listas.....	24
7 – Código em Python dos métodos de banco de dados.....	25
8 – Resultados.....	32

## LISTA DE TABELAS

1 – Categorias e quantidades.....	26
2 – Dados de treinamento.....	26
3 – Dados de detecção.....	29
4 – Resultados.....	31

## SUMÁRIO

<b>1 INTRODUÇÃO .....</b>	<b>8</b>
<b>1.1 DEFINIÇÃO DO TEMA OU PROBLEMA .....</b>	<b>9</b>
<b>1.2 DELIMITAÇÕES DO TRABALHO .....</b>	<b>9</b>
<b>1.3 OBJETIVOS.....</b>	<b>9</b>
<b>1.3.1 Objetivo Geral .....</b>	<b>9</b>
<b>1.3.2 Objetivos Específicos .....</b>	<b>9</b>
<b>1.4 JUSTIFICATIVA.....</b>	<b>10</b>
<b>2 REVISÃO BIBLIOGRÁFICA.....</b>	<b>11</b>
<b>2.1 O QUE É SPAM? .....</b>	<b>11</b>
<b>2.1.1 E-mail spam .....</b>	<b>11</b>
<b>2.1.2 Review <i>spam</i>.....</b>	<b>12</b>
<b>2.1.3 Social <i>spam</i> .....</b>	<b>13</b>
<b>2.2 O QUE É APRENDIZADO DE MÁQUINA? .....</b>	<b>13</b>
<b>2.2.1 Supervisionado .....</b>	<b>14</b>
<b>2.2.2 Não-supervisionado .....</b>	<b>14</b>
<b>2.2.3 Semi-supervisionado .....</b>	<b>15</b>
<b>2.3 O MÉTODO MAIS EFICIENTE .....</b>	<b>15</b>
<b>2.4 SQL E MYSQL.....</b>	<b>15</b>
<b>2.5 PYTHON .....</b>	<b>16</b>
<b>2.5.1 Numpy .....</b>	<b>16</b>
<b>2.5.2 Scipy.....</b>	<b>16</b>
<b>2.5.3 Scikit-learn .....</b>	<b>17</b>
<b>2.5.4 Mysql-connector .....</b>	<b>17</b>
<b>2.5.5 Flask .....</b>	<b>17</b>
<b>2.6 API .....</b>	<b>17</b>
<b>3 METODOLOGIA .....</b>	<b>18</b>
<b>3.1 PROPÓSITO .....</b>	<b>18</b>
<b>3.2 ABORDAGEM .....</b>	<b>18</b>
<b>3.3 CENÁRIO E OBJETO.....</b>	<b>18</b>
<b>3.4 PROCEDIMENTOS.....</b>	<b>18</b>

<b>4 DESENVOLVIMENTO .....</b>	<b>19</b>
<b>4.1 VISÃO .....</b>	<b>19</b>
<b>4.2 FUNCIONALIDADES.....</b>	<b>19</b>
<b>4.2.1 Aprendizado.....</b>	<b>19</b>
<b>4.2.2 Detecção .....</b>	<b>21</b>
<b>4.2.3 API .....</b>	<b>23</b>
<b>4.3 DADOS DE TREINAMENTO .....</b>	<b>26</b>
<b>4.4 DETECÇÕES .....</b>	<b>29</b>
<b>5 RESULTADOS.....</b>	<b>31</b>
<b>6 CONCLUSÃO .....</b>	<b>33</b>
<b>6.1 TRABALHOS FUTUROS.....</b>	<b>33</b>
<b>7 CRONOGRAMA .....</b>	<b>34</b>
<b>REFERÊNCIAS.....</b>	<b>35</b>

## 1 INTRODUÇÃO

O *spam* é um ente nocivo que vem causando perdas tanto pessoais quanto financeiras desde o início da internet. Para pessoas o *spam* é geralmente um incômodo. Para empresas é sempre um custo, levando em consideração o desperdício de tempo dos funcionários e o espaço utilizado no armazenamento. O que levou o *spam* a perdurar por tanto tempo é a sua capacidade de adaptação. Essa é devido aos *spammers*, pessoas ou empresas que enviam *spam*, que estão continuamente aprimorando suas técnicas de envio e escrita de *spam*. Esse melhoramento contínuo permite que os *spammers* estejam sempre um passo a frente dos sistemas de detecção. Devido a essa condição faz-se necessário a construção de novas de detecção. No momento a tecnologia que está mais atraindo a atenção tanto do mercado quanto da academia, para o uso contra o *spam*, é o aprendizado de Máquina.

O aprendizado de máquina é flexível, dados os parâmetros corretos, pode ser utilizado para a resolução de qualquer problema ou cenário. A adaptabilidade, característica que faz o *spam* tão duradouro, também é presente no aprendizado de máquina. Através do uso de suas técnicas é possível desenvolver um software que se adapte a novas situações. Assim uma vez codificado ele é capaz de resolver problemas na qual ele não foi previamente programado. Isso se dá através de treinamento. Em aprendizado de máquina, treinar o sistema significa inserir dados de exemplo, que o sistema usa para aprender novas situações. A estrutura dos dados e forma de aprendizado diferencia com base no método de aprendizado.

Dentre os métodos de aprendizado encontrados em literatura, quando o foco é o *spam*, o mais encontrado é o aprendizado supervisionado. Essa técnica tem como base o a utilização de dados rotulados para treinamento. Esse tipo de dado já possuem um valor definido e seus valores foram previamente marcados em sua estrutura. Dados rotulados dão mais trabalho para captar devido a sua necessidade de estruturação, geralmente através de trabalho, em alguma parte, manual. Portanto a rotulação dos dados é o principal impedimento dessa abordagem.

O produto na forma de API provê uma forma simples para resolução desse problema. O sistema irá receber exemplos de *spam* enviados pelo usuário. Dessa forma o usuário fica responsável pela rotulação dos dados e, portanto, pelo treinamento do sistema. Mais precisamente, esse treinamento será de sua conta.

Através da separação dos dados de exemplo de cada usuário é possível deixar o sistema personalizável. A precisão da detecção depende da personalização. O que é *spam* para um pode não ser para outro. Usa-se como exemplo uma classe de mensagens do tipo venda de ingressos do para jogos de futebol. Esse tipo de conteúdo pode ser do interesse do usuário A e não do usuário B. A API possui dois métodos principais, o primeiro é o de treinamento já explanado. O segundo é o de detecção, na qual o usuário envia textos que ele suspeita ser *spam* e o sistema retorna verdadeiro ou falso.

## 1.1 DEFINIÇÃO DO TEMA OU PROBLEMA

O aprendizado de máquina é uma ferramenta poderosa e extremamente versátil que pode ser lapidada para qualquer problema. Sendo o objetivo em questão o combate ao spam, de que maneira o aprendizado de máquina pode contribuir para a detecção de spam?

## 1.2 DELIMITAÇÕES DO TRABALHO

Esse projeto de pesquisa tem como foco a classificação de textos como *spam* ou não *spam*. Outros atributos do *spam* como informações sobre o autor e infraestrutura de envio não serão abordados.

## 1.3 OBJETIVOS

Os objetivos dividem-se em: geral e específicos.

### 1.3.1 Objetivo Geral

Encontrar soluções eficientes de detecção de spam baseadas em aprendizado de máquina.

### 1.3.2 Objetivos Específicos

- Analisar o conceito de spam e sua classificação;
- Interpretar o conceito de aprendizado de máquina e seus métodos;

- Determinar o método de aprendizado de máquina mais eficiente na detecção de spam.

#### 1.4 JUSTIFICATIVA.

“O spam é uma ameaça que está constantemente crescendo e derrubando a produtividade” (SALUNKHE et al., 2016, tradução nossa)<sup>1</sup>. Incentivos para buscar soluções para combater o *spam* não faltam. O que faltam são soluções eficazes.

O principal empecilho na batalha contra o *spam* é sua adaptabilidade. Os *spammers* estão sempre desenvolvendo novas técnicas que burlam as soluções existentes de detecção.

O aprendizado de máquina é uma tecnologia que possui esse mesmo potencial de adaptação. Porém de acordo com Crawford et al., (2015, tradução nossa) relativamente poucos estudos sobre técnicas de aprendizado de máquina para detecção de spam foram realizadas, devido ao tema ser recente<sup>2</sup>.

De acordo com estes dados justifica-se a produção de um projeto de pesquisa que vise usar o aprendizado de máquina para a detecção de *spam*.

---

<sup>1</sup> “The ever increasing menace of spam is bringing down productivity”.

<sup>2</sup> “As this domain is young, relatively few studies on machine learning techniques and review spam detection have been conducted”.

## 2 REVISÃO BIBLIOGRÁFICA

Essa seção aborda os conceitos de *spam* e aprendizado de máquina e propõem soluções que mesclam ambos os conceitos.

### 2.1 O QUE É SPAM?

O seriado britânico *Monty Python* é famoso por ter atribuído o termo *spam* a coisas indesejáveis e que são empurradas ao consumidor. No episódio 25, apresentam um apresuntado enlatado chamado '*spam*' que é servido em todas as refeições (PRATES, 2014). Esse episódio destaca a característica principal do *spam*: ser indesejável.

Com o passar do tempo o *spam* mudou de estrutura, mas sua indesejabilidade continua, de acordo com Elverton et al. (2014) sua forma atual é outra: "Um *spam* é uma mensagem de e-mail de caráter não individual e não solicitada, que é disseminada em larga escala pela rede". Aqui ele passa de comida enlatada para objeto virtual, saindo das cozinhas e invadindo a web.

O e-mail é sua forma mais conhecida, mas não é a única. Apenas e-mail não o define completamente, como diz Prates (2014):

*Spam* não está vinculado somente ao correio eletrônico comercial. O termo diz respeito à todo e qualquer tipo de mensagem indesejável e não solicitada direcionada a um número maciço de usuário da internet (PRATES, 2014, p. 33).

Em pesquisa bibliográfica, não foi localizado nenhuma classificação de *spam*, entretanto seguem os tipos que foram encontrados em fontes diversas e elencados como: e-mail *spam*, review *spam* e social *spam*.

#### 2.1.1 E-mail spam

E-mail é um método de armazenamento e encaminhamento de composição, envio, recebimento e armazenamento de mensagens em sistemas de comunicação eletrônica (JAKOBSON; RUEBEN, 2015, tradução nossa)<sup>3</sup>. Esse método também está sujeito ao *spam* na forma de e-mail *spam*.

---

<sup>3</sup> "Email is a store-and-forward method of composing, sending, receiving and storing messages over electronic communication systems".

De acordo com Renuka et al. (2015, tradução nossa) e-mail *spam* é enviar mensagens de e-mail indesejadas com conteúdo comercial para um conjunto indiscriminado de destinatários<sup>4</sup>. Esse conceito especifica que o conteúdo tem objetivo comercial. Porém Idris; Selamat (2015, tradução nossa) não faz essa classificação, definindo apenas como lixo eletrônico indesejado entregue a serviços de e-mail na internet<sup>5</sup>.

Percebe-se que ambos os métodos abordam o conceito principal do spam de indesejabilidade, podendo ser de caráter comercial ou não.

### 2.1.2 Review *spam*

*Reviews* são descrições feitas por consumidores da experiência que eles tiveram com produtos ou serviços. Como diz Kang; Park (2014, tradução nossa):

*Reviews* são caracterizados por fontes de informação produtivas providas por consumidores para monitoramento e melhoramento de níveis de satisfação do consumidor, especialmente porque eles transmitem as verdadeiras vozes dos consumidores expressando suas opiniões de forma clara (KANG; PARK, 2014, tradução nossa)<sup>6</sup>.

Esse conceito baseia-se na afirmação de que as opiniões são verdadeiras e livres de viés. Esse não é sempre o caso, pois para Mayzlin; Dover; Chevalier (2014, tradução nossa) empresas possuem incentivo para produzir *reviews* positivos para seus produtos e *reviews* negativos para seus competidores<sup>7</sup>. Esse tipo de review se define como *review spam*.

Para Crawford et al. (2015, tradução nossa) *Review spam* é uma prática na qual *spammers* manipulam e contaminam reviews para lucro ou ganhos pessoais assim deixando-os falsos, mentirosos ou enganosos<sup>8</sup>.

---

<sup>4</sup> “It is sending unwanted e-mail messages with commercial content to indiscriminate set of recipients”.

<sup>5</sup> “It is defined as an unwanted junk email delivered to services on internet mail”.

<sup>6</sup> “Customer reviews are recognized as fruitful information sources for monitoring and enhancing customer satisfaction levels, particularly as they convey the real voices of actual customers expressing relatively unambiguous opinions”.

<sup>7</sup> “Firms have an incentive to manufacture positive reviews for their own products and negative reviews for their rivals”.

<sup>8</sup> “This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews) for profit or gain”.

### 2.1.3 Social spam

O termo “*social*” desta seção refere-se as redes sociais online, que de acordo com o conceito de Villela (2016):

[...] são sistemas online projetados para permitir a interação social a partir do compartilhamento e da criação colaborativa de informação nos mais diversos formatos. Eles possibilitaram a publicação de conteúdos por qualquer pessoa [...] Elas abrangem diversas atividades que integram tecnologia, interação social, a escrita, fotos, vídeos, áudios, etc (VILLELA, 2016).

A grande popularidade das redes sociais as tornou alvo de spam, que no âmbito de redes sociais é denominado *social spam*.

Segundo Washha et al. (2017, tradução nossa) *social spam* é definido por uma mensagem de texto cujo conteúdo é besteira ou não faz sentido aparecendo em redes sociais online [...]<sup>9</sup>. A falta de sentido mencionada é com relação ao assunto do tópico ou conversa, a mensagem classificada como *spam* pode fazer sentido em outro contexto.

Ferramentas que apontam os assuntos mais populares são uma das escolhas mais populares para disseminação do *spam*, como aponta Zheng et al. (2015) “[...] Os *trending topics* são sempre abusados por spammers para publicar comentários com URLs, redirecionando todos os tipos de usuários para websites completamente não relacionados”<sup>10</sup>.

## 2.2 O QUE É APRENDIZADO DE MÁQUINA?

“Aprendizado de Máquina (AM) é caracterizado pelo desenvolvimento de técnicas que objetivam prover os softwares com a habilidade de melhorar seu desempenho [...] aprendendo através da experiência” (BRUNIALTI et al., 2015). Dessa forma os programas podem ser codificados uma vez e estarem aptos resolverem problemas que não foram pensados em suas concepções iniciais.

---

<sup>9</sup> “*Social spam is defined as a nonsentical or a gibberish text content appearing on OSNs [...]*”.

<sup>10</sup> “[...] *trending topics are always abused by spammers to publish comments with URLs, misdirecting all kinds of users to completely unrelated web-sites*”.

Segundo Pérez-Ortiz et al. (2015, tradução nossa) para identificar o algoritmo mais adequado é necessário analisar diferentes paradigmas, sendo eles supervisionado, não-supervisionado e semi-supervisionado<sup>11</sup>.

### 2.2.1 Supervisionado

No aprendizado supervisionado os dados necessariamente precisam estar rotulados, como define Zheng et al. (2015, tradução nossa)<sup>12</sup> “o aprendizado supervisionado é a tarefa de AM de inferir uma função de dados rotulados que consiste em um conjunto de exemplos de treinamento”.

Esse tipo de aprendizado baseia-se a construção de um classificador, como diz Washha et al., (2017, tradução nossa)<sup>13</sup> “métodos de aprendizado de máquina supervisionados executam treinamento para construir um classificador que possa prever corretamente as classes de novas amostras não vistas”. Quanto mais dados rotulados inseridos maior será a precisão do classificador.

### 2.2.2 Não-supervisionado

No aprendizado não-supervisionado não há a necessidade de os dados serem rotulados, como diz Zhang; Isola; Efros (2017, tradução nossa)<sup>14</sup> “o objetivo do aprendizado não-supervisionado é modelar dados brutos sem o uso de rótulos, de uma forma que produza uma representação útil”. Por não usar rótulos essa abordagem facilita a etapa de coleta de dados.

Porém esse método possui baixa performance, de acordo com Raut; Borkar, (2017, tradução nossa)<sup>15</sup> “Os algoritmos de aprendizado não supervisionados aprendem poucos recursos dos dados”.

---

<sup>11</sup> “Different machine learning paradigms are compared to identify the best performing strategies, including unsupervised, semi-supervised and supervised techniques”.

<sup>12</sup> “Supervised learning is the machine learning task of inferring a function from labeled training data that consists of a set of training examples”.

<sup>13</sup> “Classical supervised machine learning methods perform training to build a classifier that can correctly predict the classes of new unseen samples”.

<sup>14</sup> “A goal of unsupervised learning is to model raw data without the use of labels, in a manner which produces a useful representation”.

<sup>15</sup> “The unsupervised learning algorithms learn few features from the data”.

### 2.2.3 Semi-supervisionado

Para Raut; Borkar (2017, tradução nossa) O método semi-supervisionado combina o melhor de dois mundos, usa tanto técnicas de aprendizagem supervisionado e não-supervisionado<sup>16</sup>. Esse método trabalha tanto com dados rotulados e não rotulados.

A primeira vista esse paradigma parece solucionar os problemas de rotulação de dados, porem eles requerem mais poder de processamento comparados com os métodos anteriores (PÉREZ-ORTIZ et al., 2015, tradução nossa)<sup>17</sup>.

## 2.3 O MÉTODO MAIS EFICIENTE

Considerando os diferentes métodos de AM abordados nesse projeto de pesquisa, fica necessário a escolha do método que melhor desempenhe a função de detectar spam nas formas anteriormente mencionadas. O método selecionado é o supervisionado, pelos motivos explanados a seguir.

Tais métodos são os mais visados pelos pesquisadores quando o assunto é spam. “A aprendizagem supervisionada [...] é o método mais utilizado para a detecção de spam na literatura” (CRAWFORD et al., 2015, tradução nossa)<sup>18</sup>.

Os métodos de AM supervisionados são os que possuem melhor desempenho, como diz Crawford et al. (2015, tradução nossa) “[...] métodos não-supervisionados e semi-supervisionados no momento não conseguem alcançar a performance dos métodos supervisionados [...]”<sup>19</sup>.

## 2.4 SQL E MYSQL

Para Silveira; Monteiro; Souza (2010) SQL (*Structured Query Language*<sup>20</sup>) trata-se de uma linguagem para manipulação de bancos de dados. Essa linguagem é

---

<sup>16</sup> “Semi-supervised learning algorithms is a technique which combines the power of both supervised and unsupervised learning”.

<sup>17</sup> “Unlike most unsupervised methods which are relatively simple, semi-supervised ones require a higher computational load, because both labelled and unlabelled patterns are used for the training process”.

<sup>18</sup> “Supervised learning [...] is the most prevalent method used for review spam detection in the literature”.

<sup>19</sup> “[...] unsupervised and semi-supervised methods are currently unable to match the performance of supervised learning methods [...]”.

<sup>20</sup> “Linguagem de consulta estruturada”, traduzido do inglês.

utilizada pelo MySQL, que de acordo com NAIM et al., (2011, tradução nossa)<sup>21</sup>, é um sistema de gerenciamento de bancos de dados multiusuário.

## 2.5 PYTHON

Esse projeto faz uso de Python, que “é uma linguagem de programação de propósito geral” (SNYMAN; WILKE, 2018, tradução nossa)<sup>22</sup>. Por ser de propósito geral, pode ser usado para a criação de aplicações de qualquer finalidade.

Um grande atrativo que essa linguagem possui é a grande quantidade de bibliotecas que permitem a construção de aplicações sofisticadas rapidamente, como diz Oliphant (2007, tradução nossa)<sup>23</sup>. Diversas bibliotecas foram usadas nesse projeto. O motivo para o uso destas se faz na carência de conhecimento técnico em ciências como Aprendizado de máquina e cálculos matemáticos complexos e também na escolha de agilizar o processo de desenvolvimento. Essas bibliotecas são descritas nas subseções seguintes.

### 2.5.1 Numpy

De acordo com Drude et al. (2018, tradução nossa)<sup>24</sup> Numpy é uma biblioteca em Python para manipulação numérica que trabalha com grandes vetores de forma eficiente.

### 2.5.2 Scipy

Para Putza et al., (2013, tradução nossa)<sup>25</sup> Scipy é uma biblioteca em Python que contém uma grande quantidade de funções para uso científico.

---

<sup>21</sup> “MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases”.

<sup>22</sup> “Python is a general purpose computer programming language”.

<sup>23</sup> “Its large number of library modules [...] means you can quickly construct sophisticated programs”.

<sup>24</sup> “Numpy is a numerical Python library which efficiently manipulates large arrays”.

<sup>25</sup> “Scipy is a Python add-on package that provides a vast and growing library of scientific routines and functions, numerical array types, vectorized operations, logical indexing, slicing and so on”.

### 2.5.3 Scikit-learn

Segundo Pedregosa et al. (2011, tradução nossa) Scikit-learn é uma biblioteca em Python que possui uma grande variedade de algoritmos estado da arte de aprendizado de máquina<sup>26</sup>.

### 2.5.4 Mysql-connector

De acordo com Kumari; Rani; Yadav, (2017, tradução nossa)<sup>27</sup> Mysql-connector é uma biblioteca em Python que prove um driver de conexão entre a linguagem de programação Python e o banco de dados Mysql.

### 2.5.5 Flask

Outra biblioteca em Python que esse projeto utiliza é Flask, que trata-se de “um micro framework para a construção de web services” (MCLEOD, 2015, tradução nossa)<sup>28</sup>.

## 2.6 API

Para Stylos; Myers, (2007, tradução nossa)<sup>29</sup> uma API é um conjunto de código que disponível, geralmente na forma de interface, para programadores para auxílio em programação.

---

<sup>26</sup> “Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms”.

<sup>27</sup> “MySQL Python connector which is a standardized database driver provided by MySQL needed to access MySQL database from Python”.

<sup>28</sup> “Flask is a microframework for building web services”.

<sup>29</sup> “An API is a collection of existing code that other programmers can call to help accomplish programming tasks. Usually the APIs are available only in compiled form, as an interface.”

### **3 METODOLOGIA**

Essa seção descreve a metodologia utilizada. Explica o conjunto de métodos usados no desenvolvimento desse projeto.

#### **3.1 PROPÓSITO**

A metodologia desse projeto possui um propósito exploratório, devido a baixo entendimento do mesmo pelos assuntos abordados, necessitando de Levantamento bibliográfico para obter a competência necessária para a sua conclusão.

#### **3.2 ABORDAGEM**

Faz-se necessário uma abordagem qualitativa, correspondente ao aspecto intelectual dos dados usados. A reflexão para análise dos resultados é determinante para afirmar a conclusão dos objetivos.

#### **3.3 CENÁRIO E OBJETO**

Foi adotada como cenário de pesquisa o portal de notícias G1 (g1.com.br), devido a sua popularidade a nível nacional, a disponibilidade dos textos através de RSS e a diversidade do conteúdo, fazendo uso de assuntos variados.

#### **3.4 PROCEDIMENTOS**

Será adicionado no banco de dados, notícias do G1 que serviram como dados de treinamento. Diferentes notícias do mesmo portal serão usadas para testar o sistema de acordo com as classificações usadas.

## 4 DESENVOLVIMENTO

Essa seção detalha o produto resultante desse projeto de pesquisa, abordando desde sua concepção inicial a todos os detalhes técnicos envolvidos na sua produção.

### 4.1 VISÃO

Esse projeto tem como objetivo a detecção spam de forma personalizada. Para possibilitar a personalização, se faz necessário o uso do aprendizado de máquina. Objetiva-se também a gerar valor acadêmico, de forma que futuros estudos possam usa-lo como referência tanto em detecção de spam como em aprendizado de máquina.

Para atender a esses objetivos se faz necessário o desenvolvimento de um software de código aberto e que use tecnologias de fácil aprendizado e manipulação.

### 4.2 FUNCIONALIDADES

O software possui duas funcionalidades principais, o método de detecção e o de aprendizado. Para a execução desses métodos foi escolhido o modelo de API. Na Seção 4.2.1 será detalhado o método de detecção. No item 4.2.2 será descrito o método de aprendizagem. No tópico 4.2.3 o método de API será detalhado.

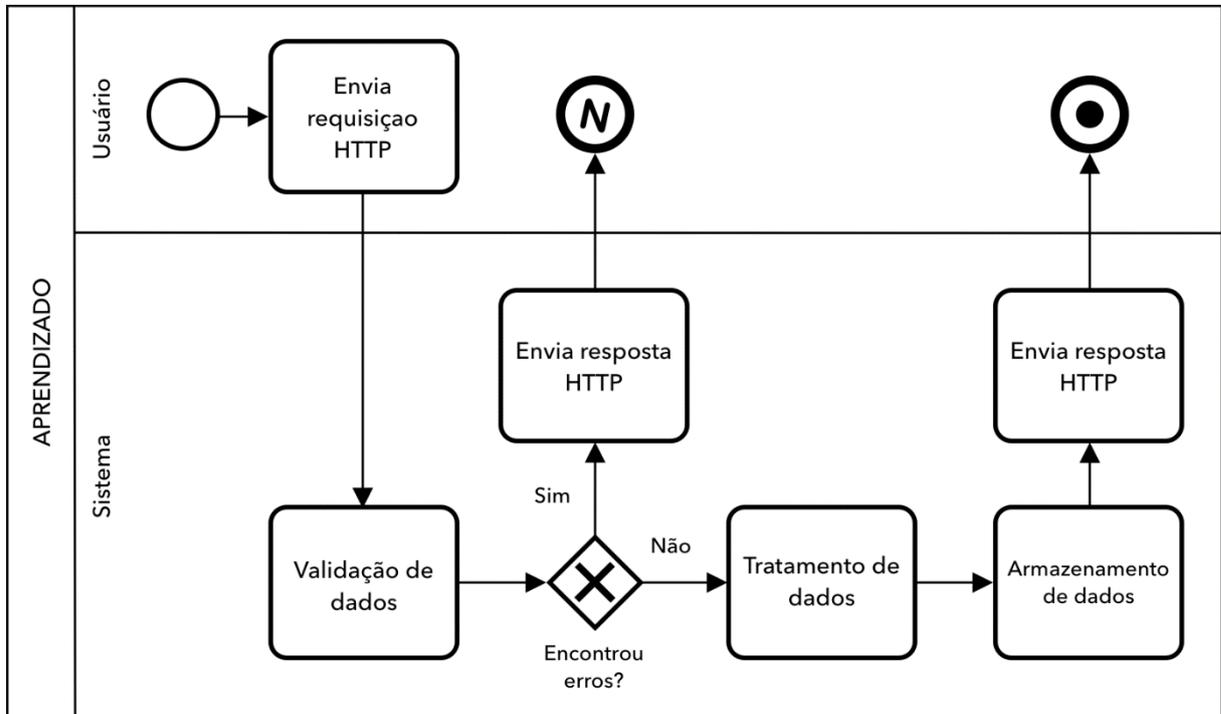
#### 4.2.1 Aprendizado

No método de aprendizado, basicamente, o usuário envia para o sistema um texto e recebe em retorno uma mensagem de envio bem-sucedido. Isso ocorre através de uma requisição HTTP que o usuário deve enviar para a API no endereço de aprendizado. O corpo da requisição deve conter dois campos: texto e classificação. O campo de texto é basicamente qualquer informação textual que não ultrapasse 5000 caracteres. O campo de classificação contem apenas uma palavra, sendo uma entre duas opções, sendo elas: “spam” e “não-spam”.

Passando por estas validações, o campo de texto é tratado de acordo com o método de tratamento de dados. Os dados são enviados para o banco de dados onde

é gerado um registro contendo o texto e a classificação. Por último é enviado ao usuário uma mensagem de confirmação no formato de uma resposta via HTTP.

**Figura 1. Fluxo de dados do método de aprendizado**



Fonte: Elaborado pelo autor.

**Figura 2. Código em Python do método de aprendizado**

```
def treinar():
    from flask import jsonify, request
    from utilidades import processar_texto
    import banco

    dados_da_requisicao = request.get_json()
    mensagem = processar_texto(dados_da_requisicao['mensagem'])
    classificacao = dados_da_requisicao['classificacao']

    banco.inserir(mensagem, classificacao)

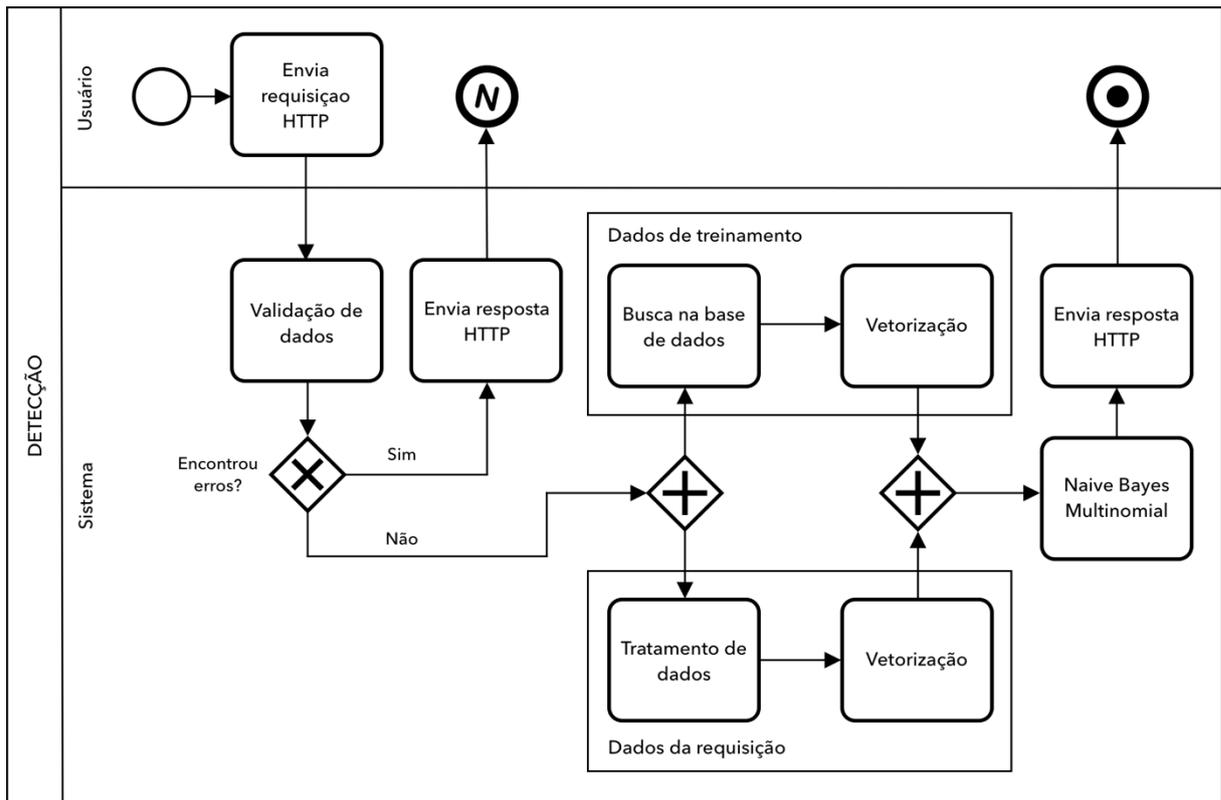
    dicionario = {
        'sucesso': True,
        'mensagem': processar_texto(dados_da_requisicao['mensagem'])
    }
    return jsonify(dicionario)
```

Fonte: Elaborado pelo autor.

### 4.2.2 Detecção

No método de detecção, sumariamente, o usuário envia um texto para o sistema e recebe em retorno uma mensagem contendo “spam” ou “não spam”. Para chegar a essa conclusão o sistema compara o texto com os exemplos em sua base de dados. Isso ocorre através de uma requisição HTTP que o usuário deve enviar para a API no endereço de detecção. O corpo da requisição deve conter apenas um campo, sendo ele o texto que precisa ser classificado como “spam” ou “não-spam”. Após o recebimento do texto ele é tratado de acordo com o método de tratamento de dados. Em seguida os dados de aprendizado são puxados do banco de dados e carregados em memória. Em seguida cada grupo de dados, o texto atual e os dados de aprendizado, são vetorizados de acordo com o método *Bag of Words*. O vetor 2 (dados de aprendizado) é usado para treinar o algoritmo Naive Bayes Multinomial. Esse algoritmo, depois de treinado, é usado para classificar o vetor 1 (texto atual). O resultado da execução do método de detecção pode ser classificado de duas formas: spam e não spam. A última etapa do método de detecção, é o envio de uma mensagem no formato de resposta HTTP contendo o resultado.

### Figura 3. Fluxo de dados do método de detecção



Fonte: Elaborado pelo autor.

**Figura 4. Código em Python do método de detecção**

```

def detectar():
    from flask import jsonify, request
    from utilidades import processar_texto
    from sklearn.feature_extraction.text import CountVectorizer
    from sklearn.naive_bayes import MultinomialNB
    import banco

    dados_da_requisicao = request.get_json()
    mensagem = processar_texto(dados_da_requisicao['mensagem'])
    dados = banco.trazer_todos_os_dados()

    vetorizador = CountVectorizer()
    valores = vetorizador.fit_transform(dados['mensagens'])

    classificador = MultinomialNB()
    classificador.fit(valores, dados['clasificacoes'])

    deteccoes = [
        mensagem
    ]
    deteccoes_valores = vetorizador.transform(detteccoes)
  
```

```

predicoes = classificador.predict(deteccoes_valores)

dicionario = {
    'mensagem': mensagem,
    'predicoes': int(predicoes[0])
}
return jsonify(dicionario)

```

Fonte: Elaborado pelo autor.

### 4.2.3 API

A estrutura escolhida para execução dos métodos é API REST. Possibilitando assim receber requisições de forma programada. Dessa forma podendo ser integrado a qualquer sistema.

### 4.2.4 O método de tratamento de dados

O método de tratamento de dados constitui-se de uma função que recebe textos e retorna-os otimizados para serem usados em algoritmos de aprendizado de máquina.

O primeiro passo é remover marcações HTML. O método considera uma marcação HTML todo elemento que começa com o caractere “<” e termina com o caractere “>”. Para detectar tais marcações se faz necessário o uso de uma expressão regular. A medida seguinte a ser tomada é deixar todo o texto em caixa baixa para evitar processamento duplo nas etapas seguintes. O próximo passo é remover caracteres especiais como vírgulas, pontos, chaves e etc. Logo após é necessário remover as palavras não significativas, isto é, palavras que não expressam sentimentos, ações ou não representam entidades.

### Figura 5. Código em Python do método de tratamento de dados

```

import re
import listas

def remover_palavras(lista_de_palavras, texto):
    texto_em_vetor = texto.split()
    texto_em_vetor_novo = []
    for palavra in texto_em_vetor:

```

```

    if palavra not in lista_de_palavras:
        texto_em_vetor_novo.append(palavra)
    texto_novo = ''.join(texto_em_vetor_novo)
    return texto_novo

def remover_html(texto):
    return re.sub('<.*?>', '', texto)

def para_caixa_baixa(texto):
    return texto.lower()

def deixar_apenas_palavras(texto):
    texto_em_vetor = texto.split()
    texto_em_vetor_novo = []
    for palavra in texto_em_vetor:
        palavra_nova = ""
        for caracter in palavra:
            if caracter.isalpha():
                palavra_nova = palavra_nova + caracter
        if len(palavra_nova) > 0:
            texto_em_vetor_novo.append(palavra_nova)
    texto_novo = ''.join(texto_em_vetor_novo)
    return texto_novo

def remover_palavras_nao_relevantes(texto):
    lista_de_palavras = listas.palavras_nao_relevantes
    return remover_palavras(lista_de_palavras, texto)

def processar_texto(texto):
    passo_1 = remover_html(texto)
    passo_2 = para_caixa_baixa(passo_1)
    passo_3 = deixar_apenas_palavras(passo_2)
    texto_pronto = remover_palavras_nao_relevantes(passo_3)
    return texto_pronto

```

Fonte: Elaborado pelo autor.

### Figura 6. Código em Python utilizado para gerar listas

```

palavras_nao_relevantes = [
    'o','e','é','os','a',
    'as','um','uns','uma','umas',
    'a','ao','aos','à','às','de',
    'do','dos','da','das','dum',
    'duns','duma','dumas','em','no',

```

```

'nos','na','nas','num','nuns',
'numa','numas','por','per','pelo',
'pelos','pela','pelas','com','que',
'tem','como','foi','mas','me',
'pra','ai','se','para','isso',
'essa','esse','ter','ser','vai',
'foi','dar'
]

```

Fonte: Elaborado pelo autor.

#### 4.2.5 Os métodos de banco de dados

Figura 7. Código em Python dos métodos de banco de dados

```

import mysql.connector

banco = mysql.connector.connect(
    host = 'db',
    user = 'root',
    passwd = 'password',
    database = 'banco'
)
banco_cursor = banco.cursor()

def inserir(mensagem, classificacao):
    sql = 'INSERT INTO dados (mensagem, classificacao) VALUES (%s, %s)'
    valores = (
        mensagem,
        classificacao
    )
    banco_cursor.execute(sql, valores)
    banco_cursor.commit()

def trazer_todos_os_dados():
    sql = 'SELECT * from dados'
    banco_cursor.execute(sql)
    resultado = banco_cursor.fetchall()
    mensagens = []
    classificacoes = []
    for item in resultado:
        mensagens.append(item[1])
        classificacoes.append(item[2])
    dicionario = {

```

```
'mensagens': mensagens,
'clasificacoes': classificacoes
}
return dicionario
```

Fonte: Elaborado pelo autor.

### 4.3 DADOS DE TREINAMENTO

Para ensinar o sistema, foi determinada a quantidade de 24 itens a serem inseridos no banco através do método de treinamento. Esses itens se fazem de notícias, metade delas (12) são oriundas da categoria música, classificadas dessa forma pelo próprio G1. A tabela 1 mostra a relação das categorias com as quantidades e também as URLs usadas para obtenção dos mesmos.

**Tabela 1. Categorias e quantidades**

CATEGORIA	QUANTIDADE	URL
Música	12	<a href="https://g1.globo.com/pop-arte/musica">https://g1.globo.com/pop-arte/musica</a>
Tecnologia	3	<a href="https://g1.globo.com/economia/tecnologia">https://g1.globo.com/economia/tecnologia</a>
Política	3	<a href="https://g1.globo.com/politica">https://g1.globo.com/politica</a>
Educação	3	<a href="https://g1.globo.com/educacao">https://g1.globo.com/educacao</a>
Carros	3	<a href="https://g1.globo.com/carros">https://g1.globo.com/carros</a>

Fonte: Elaborado pelo autor.

A tabela 2 mostra a relação das categorias e suas classificações usadas (spam ou não spam) e os seus respectivos endereços eletrônicos (URL).

**Tabela 2. Dados de treinamento**

NÚMERO	CATEGORIA	CLASSIFICAÇÃO	URL DO G1	TEXTO USADO
1	Música	Não spam	<a href="https://g1.globo.com/pb/parai-ba/sao-joao/2018/ao-vivo/sao-joao-2018-de-campina-grande.ghtml">https://g1.globo.com/pb/parai-ba/sao-joao/2018/ao-vivo/sao-joao-2018-de-campina-grande.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/1.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/1.txt</a>

2	Música	Não spam	<a href="https://g1.globo.com/pe/pernambuco/musica/noticia/cantores-se-mobilizam-para-transformar-o-forro-em-patrimonio-imaterial-do-brasil.ghtml">https://g1.globo.com/pe/pernambuco/musica/noticia/cantores-se-mobilizam-para-transformar-o-forro-em-patrimonio-imaterial-do-brasil.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/2.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/2.txt</a>
3	Música	Não spam	<a href="https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/aos-64-anos-professor-aposentado-lanca-album-autoral-de-forro-durante-o-sao-joao.ghtml">https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/aos-64-anos-professor-aposentado-lanca-album-autoral-de-forro-durante-o-sao-joao.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/3.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/3.txt</a>
4	Música	Não spam	<a href="https://g1.globo.com/pop-arte/noticia/paul-mccartney-canta-musica-de-novo-album-em-entrevista-e-video-viraliza.ghtml">https://g1.globo.com/pop-arte/noticia/paul-mccartney-canta-musica-de-novo-album-em-entrevista-e-video-viraliza.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/4.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/4.txt</a>
5	Música	Não spam	<a href="https://g1.globo.com/musica/noticia/vinnie-paul-baterista-do-pantera-morre-aos-54-anos.ghtml">https://g1.globo.com/musica/noticia/vinnie-paul-baterista-do-pantera-morre-aos-54-anos.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/5.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/5.txt</a>
6	Música	Não spam	<a href="https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/fagner-faz-show-no-sao-joao-2018-de-campina-grande-neste-sabado-23.ghtml">https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/fagner-faz-show-no-sao-joao-2018-de-campina-grande-neste-sabado-23.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/6.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/6.txt</a>
7	Música	Não spam	<a href="https://g1.globo.com/ce/ceara/noticia/cantor-wesley-safadao-passa-mal-e-e-hospitalizado-em-fortaleza.ghtml">https://g1.globo.com/ce/ceara/noticia/cantor-wesley-safadao-passa-mal-e-e-hospitalizado-em-fortaleza.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/7.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/7.txt</a>
8	Música	Não spam	<a href="https://g1.globo.com/pop-arte/noticia/beyonce-e-jay-z-lancam-album-surpresa.ghtml">https://g1.globo.com/pop-arte/noticia/beyonce-e-jay-z-lancam-album-surpresa.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/8.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/8.txt</a>
9	Música	Não spam	<a href="https://g1.globo.com/musica/noticia/aldair-playboy-quem-e-o-cantor-que-levou-amor-falso-ao-1o-lugar-no-brasil-e-ate-ao-treino-da-selecao-da-russia.ghtml">https://g1.globo.com/musica/noticia/aldair-playboy-quem-e-o-cantor-que-levou-amor-falso-ao-1o-lugar-no-brasil-e-ate-ao-treino-da-selecao-da-russia.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/9.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/9.txt</a>
10	Música	Não spam	<a href="https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/quero-mostrar-que-campina-">https://g1.globo.com/pb/paraba/sao-joao/2018/noticia/quero-mostrar-que-campina-</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/10.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/10.txt</a>

			grande-tem-o-maior-sao-joao-diz-luan-santana-no-parque-do-povo.ghtml	
11	Música	Não spam	<a href="https://g1.globo.com/pb/parai-ba/sao-joao/2018/noticia/abertura-do-sao-joao-2018-de-campina-grande-tem-mano-walter-santanna-e-show-pirotecnico.ghtml">https://g1.globo.com/pb/parai-ba/sao-joao/2018/noticia/abertura-do-sao-joao-2018-de-campina-grande-tem-mano-walter-santanna-e-show-pirotecnico.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/11.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/11.txt</a>
12	Música	Não spam	<a href="https://g1.globo.com/musica/noticia/gravadora-de-princes-lanca-album-postumo-por-seu-60o-aniversario.ghtml">https://g1.globo.com/musica/noticia/gravadora-de-princes-lanca-album-postumo-por-seu-60o-aniversario.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/12.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/12.txt</a>
13	Tecnologia	Spam	<a href="https://g1.globo.com/carros/noticia/elon-musk-diz-que-fabrica-da-tesla-sofreu-sabotagem-de-funcionario.ghtml">https://g1.globo.com/carros/noticia/elon-musk-diz-que-fabrica-da-tesla-sofreu-sabotagem-de-funcionario.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/13.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/13.txt</a>
14	Tecnologia	Spam	<a href="http://g1.globo.com/tecnologia/blog/seguranca-digital/post/golpistas-distribuem-extensoes-falsas-do-chrome-no-facebook.html">http://g1.globo.com/tecnologia/blog/seguranca-digital/post/golpistas-distribuem-extensoes-falsas-do-chrome-no-facebook.html</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/14.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/14.txt</a>
15	Tecnologia	Spam	<a href="http://g1.globo.com/tecnologia/blog/seguranca-digital/post/certificado-digital-do-banco-inter-e-revogado-apos-chave-vazar-na-web.html">http://g1.globo.com/tecnologia/blog/seguranca-digital/post/certificado-digital-do-banco-inter-e-revogado-apos-chave-vazar-na-web.html</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/15.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/15.txt</a>
16	Política	Spam	<a href="https://g1.globo.com/politica/noticia/2019/06/22/bolsonaro-critica-projeto-aprovado-pelo-congresso-querem-me-deixar-como-rainha-da-inglaterra.ghtml">https://g1.globo.com/politica/noticia/2019/06/22/bolsonaro-critica-projeto-aprovado-pelo-congresso-querem-me-deixar-como-rainha-da-inglaterra.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/16.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/16.txt</a>
17	Política	Spam	<a href="https://g1.globo.com/politica/noticia/2019/06/21/pedido-de-liberdade-de-lula-nao-se-baseia-em-reportagens-de-site-diz-defesa-apos-parecer-da-pgr.ghtml">https://g1.globo.com/politica/noticia/2019/06/21/pedido-de-liberdade-de-lula-nao-se-baseia-em-reportagens-de-site-diz-defesa-apos-parecer-da-pgr.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/17.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/17.txt</a>
18	Política	Spam	<a href="https://g1.globo.com/politica/noticia/2019/06/21/bolsonaro-discursara-em-sessao-tematica-sobre-tecnologia-e-">https://g1.globo.com/politica/noticia/2019/06/21/bolsonaro-discursara-em-sessao-tematica-sobre-tecnologia-e-</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/18.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/18.txt</a>

			inovacao-na-cupula-do-g20-no-japao.ghtml	
19	Educação	Spam	<a href="https://g1.globo.com/educacao/noticia/2019/06/14/paralisacoes-escolas.ghtml">https://g1.globo.com/educacao/noticia/2019/06/14/paralisacoes-escolas.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/19.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/19.txt</a>
20	Educação	Spam	<a href="https://g1.globo.com/educacao/noticia/2019/06/14/prounido-2o-semester-encerrainscricoes-nesta-sexta-14.ghtml">https://g1.globo.com/educacao/noticia/2019/06/14/prounido-2o-semester-encerrainscricoes-nesta-sexta-14.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/20.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/20.txt</a>
21	Educação	Spam	<a href="https://g1.globo.com/rj/rio-de-janeiro/noticia/2019/06/13/inscricoes-para-temporarios-da-faetec-terminam-nesta-sexta.ghtml">https://g1.globo.com/rj/rio-de-janeiro/noticia/2019/06/13/inscricoes-para-temporarios-da-faetec-terminam-nesta-sexta.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/21.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/21.txt</a>
22	Carros	Spam	<a href="https://g1.globo.com/economia/tecnologia/noticia/2019/06/12/uber-testa-entrega-de-comida-com-drones-e-apresenta-novo-veiculo-autonomo.ghtml">https://g1.globo.com/economia/tecnologia/noticia/2019/06/12/uber-testa-entrega-de-comida-com-drones-e-apresenta-novo-veiculo-autonomo.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/22.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/22.txt</a>
23	Carros	Spam	<a href="https://g1.globo.com/sp/sao-paulo/noticia/2019/06/12/patinete-eletrico-regras-para-rodar-em-sao-paulo.ghtml">https://g1.globo.com/sp/sao-paulo/noticia/2019/06/12/patinete-eletrico-regras-para-rodar-em-sao-paulo.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/23.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/23.txt</a>
24	Carros	Spam	<a href="https://g1.globo.com/carros/noticia/2019/06/12/prioridade-da-renault-e-fortalecer-alianca-com-nissan-diz-presidente-do-conselho.ghtml">https://g1.globo.com/carros/noticia/2019/06/12/prioridade-da-renault-e-fortalecer-alianca-com-nissan-diz-presidente-do-conselho.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/24.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/treinamento/texto/24.txt</a>

Fonte: Elaborado pelo autor.

#### 4.4 DETECÇÕES

Para testar o sistema, foi determinado que oito (8) usos do método de detecção seriam realizados. A tabela 3 relaciona os dados usados, categorias e URLs.

**Tabela 3. Dados de detecção**

NÚMERO	CATEGORIA	URL DO G1	TEXTO USADO
--------	-----------	-----------	-------------

1	Música	<a href="https://g1.globo.com/musica/blog/mauro-ferreira/post/2018/04/06/de-olho-no-lance-rincon-bate-bolao-com-rael-e-conka-em-samba-rap-sobre-futebol.ghtml">https://g1.globo.com/musica/blog/mauro-ferreira/post/2018/04/06/de-olho-no-lance-rincon-bate-bolao-com-rael-e-conka-em-samba-rap-sobre-futebol.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/1.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/1.txt</a>
2	Música	<a href="https://g1.globo.com/musica/blog/mauro-ferreira/post/2018/04/06/zeze-motta-faz-show-continuar-e-lanca-disco-de-samba-em-que-canta-com-arlindo.ghtml">https://g1.globo.com/musica/blog/mauro-ferreira/post/2018/04/06/zeze-motta-faz-show-continuar-e-lanca-disco-de-samba-em-que-canta-com-arlindo.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/2.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/2.txt</a>
3	Música	<a href="https://g1.globo.com/musica/noticia/pablo-vittar-figurino-da-cantora-no-clipe-de-indestrutivel-sera-leiloado-para-ajudar-projeto-lgbt.ghtml">https://g1.globo.com/musica/noticia/pablo-vittar-figurino-da-cantora-no-clipe-de-indestrutivel-sera-leiloado-para-ajudar-projeto-lgbt.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/3.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/3.txt</a>
4	Música	<a href="https://g1.globo.com/economia/tecnologia/noticia/spotify-compra-empresa-de-licenciamento-de-musica-cover-para-conter-riscos-de-direitos-autorais.ghtml">https://g1.globo.com/economia/tecnologia/noticia/spotify-compra-empresa-de-licenciamento-de-musica-cover-para-conter-riscos-de-direitos-autorais.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/4.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/4.txt</a>
5	Tecnologia	<a href="http://g1.globo.com/tecnologia/blog/тира-duvidas-de-tecnologia/post/facebook-e-possivel-ter-conta-hackeada-so-portal-aceito-uma-solicitacao-de-amizade.html">http://g1.globo.com/tecnologia/blog/тира-duvidas-de-tecnologia/post/facebook-e-possivel-ter-conta-hackeada-so-portal-aceito-uma-solicitacao-de-amizade.html</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/5.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/5.txt</a>
6	Política	<a href="https://g1.globo.com/rj/rio-de-janeiro/noticia/2019/06/19/stj-manda-soltar-homem-que-ficou-preso-7-meses-acusado-de-dar-auxilio-moral-a-criminosos.ghtml">https://g1.globo.com/rj/rio-de-janeiro/noticia/2019/06/19/stj-manda-soltar-homem-que-ficou-preso-7-meses-acusado-de-dar-auxilio-moral-a-criminosos.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/6.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/6.txt</a>
7	Educação	<a href="https://g1.globo.com/educacao/noticia/2019/06/10/sisu-2019-divulga-lista-de-aprovados-na-edicao-do-2o-semester.ghtml">https://g1.globo.com/educacao/noticia/2019/06/10/sisu-2019-divulga-lista-de-aprovados-na-edicao-do-2o-semester.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/7.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/7.txt</a>
8	Carros	<a href="https://g1.globo.com/carros/motos/noticia/2019/06/10/kawasaki-versys-1000-2020-chega-renovada-ao-brasil-por-r-55490.ghtml">https://g1.globo.com/carros/motos/noticia/2019/06/10/kawasaki-versys-1000-2020-chega-renovada-ao-brasil-por-r-55490.ghtml</a>	<a href="https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/8.txt">https://bitbucket.org/marinhod/tcc-dados/src/master/deteccao/8.txt</a>

Fonte: Elaborado pelo autor.

## 5 RESULTADOS

A revisão bibliográfica, as 24 notícias, retiradas do g1.com.br, inseridas no sistema como dados de treinamento e o próprio sistema, no formato de API, serviram de fonte para o estudo de uma API para detecção de spam utilizando aprendizado de máquina.

Após o desenvolvimento do sistema, chegou o momento de alimentar a base de dados com os dados de treinamento. Para isso é utilizado o método de treinamento. A quantidade de inserções escolhida foi vinte e quatro (24). Metade a ser classificado como spam e a outra metade como não spam. É uma quantidade baixa, porem com propósito, se o sistema apresentar o funcionamento esperado com uma baixa quantidade de dados de treinamento, significa que o usuário já vai ter um bom resultado logo no início do uso, considerando que faz parte do proposito desse projeto ser customizado para cada usuário.

Após inseridos os dados com suas devidas classificações, chega o momento de testar o sistema utilizando o método de detecção. A quantidade de detecções escolhida foi oito (8). O comportamento esperado é que metade (4) seja classificado como spam e a outra metade (4) seja classificado como não spam.

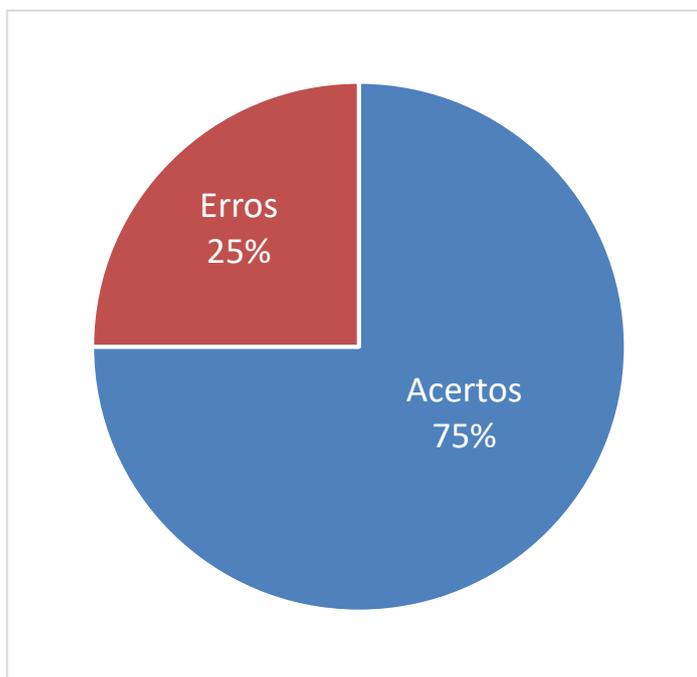
**Tabela 4. Resultados**

NÚMERO	CLASSIFICAÇÃO ESPERADA	CLASSIFICAÇÃO OBTIDA	OBTIDA = ESPERADA
1	Não Spam	Não Spam	Sim
2	Não Spam	Não Spam	Sim
3	Não Spam	Não Spam	Sim
4	Não Spam	Não Spam	Sim
5	Spam	Não Spam	Não
6	Spam	Spam	Sim
7	Spam	Spam	Sim
8	Spam	Não Spam	Não

Fonte: Elaborado pelo autor.

Os resultados obtidos nas detecções, como mostra na tabela 4, são de seis (6) acertos e dois (2) erros. Esperava-se que as quatro (4) detecções da categoria música retornassem como não spam e as quatro (4) detecções das outras categorias retornassem como spam. A figura 8 faz uma apresentação visual dos resultados.

**Figura 8. Resultados**



Fonte: Elaborado pelo autor.

Com base na média aritmética, o resultado obtido se traduz em 75% de acerto. Assim são apresentados os valores na figura 8, sendo eles 75% de acertos e 25% de erros.

## 6 CONCLUSÃO

Esse trabalho apresentou estudos sobre AM e spam, também desenvolveu um sistema na forma de API, na tentativa de encontrar maneiras de utilizar o AM para detecção de spam. Houve dificuldades iniciais nas questões relacionadas a AM, pois tratava-se de uma área de estudo completamente nova para o autor.

As expectativas eram de alcançar todos os objetivos propostos por esse projeto, e de fato todos eles foram atingidos. O conceito de spam e suas classificações foram devidamente analisados. O conceito de aprendizado de máquina e seus métodos foram propriamente interpretados. E uma solução eficiente para detecção de spam baseado em AM foi encontrada. Essa solução trata-se do sistema desenvolvido nesse projeto. Os resultados obtidos são positivos, o que faz com que essa solução seja considerada eficiente.

### 6.1 TRABALHOS FUTUROS

A biblioteca *Scikit-learn* possui diversos algoritmos além do *Multinomial Naive Bayes*, utilizado nesse projeto, quem podem ser testados e comparados para uso na detecção de spam. Futuras implementações podem utilizar a API implementada nesse projeto para testar os outros algoritmos.

Indo além do spam, a API proposta também pode ser usada, com pequenas modificações, para classificar qualquer categoria. A implementação atual fixa as categorias em “spam” e “não spam”, porem teoricamente qualquer texto pode ser usado para dar nome a classificação. Dessa forma pode ser desenvolvido um classificador genérico de notícias, ou qualquer outro formato textual.

## 7 CRONOGRAMA

ATIVIDADES	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.
Escolha do assunto do projeto	X					
Elaboração da estrutura do projeto	X					
Seleção e leitura das obras para elaboração do projeto		X				
Elaboração dos objetivos, delimitação do tema, definição do problema, etc.		X				
Elaboração da pesquisa bibliográfica e documental do projeto			X			
Coleta de dados				X		
Tratamento dos dados				X		
Revisão final do texto e elaboração da introdução e conclusão					X	
Entrega do Projeto						X

## REFERÊNCIAS

- BRUNIALTI, L. F. et al. Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática. **XI Brazilian Symposium on Information System**, [s. l.], p. 203–210, 2015.
- CRAWFORD, M. et al. Survey of review spam detection using machine learning techniques. **Journal of Big Data**, [s. l.], p. 2–23, 2015.
- DRUDE, L. et al. NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing. **Speech Communication**, [s. l.], v. Speech Communication; 13th ITG-Symposium, p. 5, 2018.
- ELVERTON, F. et al. SpamBands: uma metodologia para identificação de fontes de spam agindo de forma orquestrada. **XIV Simpósio Brasileiro em Segurança e de Sistemas Computacionais**, [s. l.], p. 1–14, 2014.
- IDRIS, I.; SELAMAT, A. A Swarm Negative Selection Algorithm for Email Spam Detection. **Journal of Computer Engineering & Information Technology**, [s. l.], p. 1–4, 2015.
- JAKOBSON, G.; RUEBEN, S. Method and system for processing and displaying email thread information. [s. l.], p. 29, 2015.
- KANG, D.; PARK, Y. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. **Expert Systems with Applications**, [s. l.], v. 41, p. 1041–1050, 2014.
- KUMARI, S.; RANI, K. S.; YADAV, M. Database Connection Technology. **International Journal of Advanced Research in Computer Science**, [s. l.], v. 8, n. 5, p. 2138–2143, 2017.
- MAYZLIN, D.; DOVER, Y.; CHEVALIER, J. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. **American Economic Review**, [s. l.], v. 104, n. 8, p. 2421–2455, 2014.
- MCLEOD, C. Framework for Distributed Deep Learning Layer Design in Python. **Cornell University**, [s. l.], p. 8, 2015.

NAIM, N. F. et al. MySQL Database for Storage of Fingerprint Data. **International Conference on Modelling and Simulation**, [s. l.], v. 13, p. 293–298, 2011.

OLIPHANT, T. E. Python for Scientific Computing. **Computing in Science and Engineering**, [s. l.], v. 9, n. 3, p. 10–20, 2007.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learnin**, [s. l.], v. 12, p. 6, 2011.

PÉREZ-ORTIZ, M. et al. A semi-supervised system for weed mapping in sunflower crops usingunmanned aerial vehicles and a crop row detection method. **Applied Soft Computing**, [s. l.], v. 37, p. 533–544, 2015.

PRATES, C. C. Privacidade e Intimidade na Internet: A Legalidade dos Cookies e Spam. **FMU Direito**, [s. l.], v. 42, p. 29–45, 2014.

PUTZA, A. et al. Introducing OpenPNM: An Open Source Pore Network Modeling Software Package. **ECS - The Electrochemical Society**, [s. l.], v. 58, n. 1, p. 79–86, 2013.

RAUT, P. P.; BORKAR, N. R. Machine Learning Algorithms:Trends, Perspectives and Prospects. **International Journal of Engineering Science and Computing**, [s. l.], v. 7, n. 3, p. 4884–4891, 2017.

RENUKA, D. K.; VISALAKSHI, P.; SANKAR, T. Improving E-Mail Spam Classification using Ant Colony Optimization Algorithm. **International Journal of Computer Applications**, [s. l.], p. 22–26, 2015.

SALUNKHE, N. et al. Spam Email Classification & Blocking. **Journal of Advance Research in Computer Science and Management Studies**, [s. l.], v. 4, n. 1, p. 109–115, 2016.

SILVEIRA, M. C. Da; MONTEIRO, J.´ M.; SOUZA, J. T. De. Um Ambiente de m-Learning para Ensino da Linguagem SQL. **BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION**, [s. l.], p. 1521–2266, 2010.

SNYMAN, J. A.; WILKE, D. N. Pratical computational optimization using python. **Springer International Publishing AG**, [s. l.], v. 133, p. 311–340, 2018.

STYLOS, J.; MYERS, B. Mapping the Space of API Design Decisions. **IEEE Symposium on Visual Languages and Human-Centric Computing**, [s. l.], p. 50–57, 2007.

VILLELA, F. F. Chega de bullying! A produção e difusão de material didático voltado para a prevenção do bullying em escolas de meio rural. **Revista Eletrônica de Ciências da Educação**, [s. l.], v. 15, n. 2, p. 19, 2016.

WASHHA, M. et al. A Topic-Based Hidden Markov Model for Real-Time Spam Tweets Filtering. **Procedia Computer Science**, [s. l.], v. 112, p. 833–843, 2017.

ZHANG, R.; ISOLA, P.; EFROS, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. **IEEE Xplore**, [s. l.], p. 1058–01067, 2017.

ZHENG, X. et al. Detecting Spammers on social networks. **Neurocomputing**, [s. l.], v. 159, p. 27–34, 2015.