



FACULDADE DE TECNOLOGIA DO IPIRANGA
CURSO DE ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Gianluca Silva Campana Ferreira
Guilherme Vieira Sousa
Matheus Barnabé Pereira

**Detector de Fake News em língua portuguesa
via aprendizado de máquina**

São Paulo
2022

Gianluca Silva Campana Ferreira

Guilherme Vieira Sousa

Matheus Barnabé Pereira

**Detector de Fake News em língua Portuguesa
via aprendizado de máquina**

Trabalho de Conclusão de Curso
apresentado à Faculdade de Tecnologia
do Ipiranga, como requisito parcial para
a obtenção do grau de Tecnólogo em
Análise e Desenvolvimento de
Sistemas.

Data de aprovação:

Banca examinadora:

Prof.(título) nome do professor

Presidente da Banca

Prof.(título) nome do professor

Professor Convidado

Prof. MsC Carlos Eduardo Dantas de Menezes

Professor Orientador

São Paulo

2022

Dedico este trabalho a Deus que é primordial para minha vida, aos meus pais juntamente da minha namorada que sempre me incentivaram e me ensinaram a correr atrás dos meus sonhos e objetivos, além de serem pessoas fundamentais na minha vida, demais familiares, irmã, vó, tios e tias por todo o apoio.

Matheus Barnabé Pereira

Agradecimentos

Agradecemos a todo o corpo docente da FATEC Ipiranga, por terem compartilhado seus conhecimentos no decorrer do curso que hoje são experiências e aprendizados que serão levadas por toda nossa vida.

Agradecimento aos amigos, colegas de sala e a nós que tivemos apoio de um ao outro para concluirmos este trabalho.

Agradecimento ao nosso orientador Carlos Eduardo Dantas de Menezes que acompanhou o nascimento do tema lá em 2021 e desde então vem orientando e apoiando nosso grupo.

É... palavras comovem, as atitudes convencem.” Sócrates

Resumo

Com a popularização da internet, redes sociais, aplicativos de mensagens, facilitou muito a disseminação de notícias, e com esses fatores as notícias falsas que sempre existiram, vieram com mais força e suas consequências se agravaram muito mais. As *Fake News* são um verdadeiro desserviço social, podendo notícias falsas desde coisas como terra plana até interferência em eleições presenciais. Esse trabalho tem o objetivo de identificar notícias falsas e verdadeira com criação de uma página web, e via modelos preditivos com as técnicas *TF-IDF* e *Bag of Words*, além da utilização do algoritmo de aprendizado máquina via regressão logística possibilitando com que o software faça detecção de possíveis notícias falsas.

Com base em testes foram usados três algoritmos de predição SVM (acurácia de 96,46%), Regressão Logística (acurácia de 95,9%) e MLPClassifier (acurácia de 96,4%), todos eles com a técnica TF-IDF. Foram usados os 3 melhores resultados para que a predição na página tenha uma maior veracidade.

Para a criação da página, será utilizado o *framework Flask* da linguagem de programação *Python*, *CSS* para adicionar estilo a aplicação e o banco de dados relacional *SQLServer*, para gravar os usuários e suas notícias identificadas.

Palavras-chave: Fake News. Notícias. Aprendizado de máquina. Software. Regressão logística. *TF-IDF*. *Bag of Words*. *Flask*. *Python*. *CSS*. *MySQL*.

Abstract

With the popularization of the internet, social networks, messaging applications, it greatly facilitated the dissemination of news, and with these factors such as false news that always existed, they came with more force and their consequences were much worse. Fake News is a real social disservice, and can feature fake news from things like flat earth to interference in face-to-face elections. This work aims to identify true and false news with the creation of a web page, and via predictive models with the TF-IDF and Bag of Words techniques, in addition to the use of the machine learning algorithm via logistic regression, allowing the software to detect possible fake news.

Based on tests, three SVM prediction algorithms (96.46% accuracy), Logistic Regression (95.9% accuracy) and MLPClassifier (96.4% accuracy) were used, all of them with the TF- technique IDF. The top 3 results were used so that the prediction on the page has greater accuracy.

To create the page, the Flask framework of the Python programming language will be used, CSS to add style to the application and the relational database SQLServer, to record users and their identified news.

Keywords: News. Machine learning. Logistic regression. Python. *TF-IDF*. *Bag of Words*. *Flask*. *Python*. *CSS*. *MySQL*

Sumário

1	INTRODUÇÃO.....	15
1.1	OBJETIVO.....	16
1.2	OBJETIVO GERAL.....	16
1.3	JUSTIFICATIVA E MOTIVAÇÕES.....	16
1.4	ORGANIZAÇÃO DO TRABALHO.....	16
2	PROPOSTA DE DESENVOLVIMENTO DO PROJETO.....	18
2.1	O QUE É FAKE NEWS.....	18
2.1.1	<i>A presença de Fake News em redes sociais.....</i>	<i>18</i>
2.1.2	<i>Consequências das Fake News.....</i>	<i>21</i>
2.1.3	<i>Como se prevenir das Fake News.....</i>	<i>22</i>
2.2	MINERAÇÃO DE DADOS.....	22
2.2.1	<i>Tarefas e Técnicas de Mineração de Dados.....</i>	<i>23</i>
2.2.2	<i>Processamento de Linguagem Natural.....</i>	<i>23</i>
2.2.3	<i>Feature Extraction.....</i>	<i>24</i>
2.2.4	<i>Stopwords.....</i>	<i>24</i>
2.2.5	<i>Tokenização.....</i>	<i>25</i>
2.2.6	<i>Stemming.....</i>	<i>25</i>
2.2.7	<i>TF-IDF.....</i>	<i>26</i>
2.2.8	<i>Bag of words.....</i>	<i>26</i>
2.3	INTELIGÊNCIA ARTIFICIAL.....	27
2.3.1	<i>Machine Learning.....</i>	<i>27</i>
2.3.2	<i>Supervisionado.....</i>	<i>27</i>
2.3.3	<i>Não supervisionado.....</i>	<i>27</i>
2.3.4	<i>Por reforço.....</i>	<i>28</i>
2.3.5	<i>Deep Learning.....</i>	<i>28</i>
2.4	ALGORITMOS DE PREDIÇÃO.....	28
2.4.1	<i>Regressão Logística.....</i>	<i>29</i>

2.4.2	SVM.....	30
2.4.3	Naive Bayes.....	31
2.4.4	MLPClassifier.....	32
2.5	RECURSOS.....	33
2.5.1	Python.....	33
2.5.2	Flask.....	34
2.5.3	JavaScript.....	34
2.5.4	Banco de dados Relacional.....	35
2.5.5	MySQL.....	35
2.5.6	HTML.....	36
2.5.7	CSS.....	36
2.5.8	Bootstrap.....	37
2.5.9	MVC.....	37
3	REQUISITOS DO SISTEMA DE SOFTWARE.....	38
3.1	REQUISITOS FUNCIONAIS.....	38
3.2	REQUISITOS NÃO-FUNCIONAIS.....	39
3.3	MODELAGEM FUNCIONAL.....	40
3.3.1	Diagrama de Caso de Uso.....	40
3.3.2	Atores.....	41
3.3.3	Especificação do Caso de Uso.....	41
4	ANÁLISE.....	46
4.1	DIAGRAMA DE CLASSES DE ANÁLISE (VISÃO DE NEGÓCIO).....	46
5	PROJETO.....	47
5.1	ARQUITETURA DO SISTEMA.....	47
5.2	DIAGRAMA DE CLASSES DE PROJETO POR CASO DE USO.....	47
5.3	DIAGRAMA DE SEQUÊNCIAS.....	51
5.4	DIAGRAMA DE ATIVIDADES.....	57

5.5	DIAGRAMA DE ESTADOS.....	61
6	TESTES.....	62
6.1	PLANO DE TESTES.....	62
6.1.1	<i>Escopo.....</i>	62
6.2	ITENS-ALVO DOS TESTES.....	62
6.3	RESUMO DOS TESTES PLANEJADOS.....	63
6.3.1	<i>Resumo das inclusões dos testes.....</i>	63
6.3.2	<i>Resumo dos outros candidatos a possível inclusão.....</i>	63
6.3.3	<i>Necessidades Ambientais.....</i>	63
6.3.4	<i>Elementos de softwares básicos do ambiente de teste.....</i>	64
6.4	RESPONSABILIDADES, PERFIL DA EQUIPE E NECESSIDADES DE TREINAMENTO.....	64
6.4.1	<i>Hardware básico do sistema.....</i>	64
7	RESULTADOS.....	66
7.1	DATASET.....	66
7.2	PRÉ-PROCESSAMENTO.....	67
7.3	PALAVRAS MAIS FREQUENTES NO DATASET POR INTEIRO.....	67
7.3.1	<i>Análise das notícias falsas.....</i>	68
7.3.2	<i>Análise das notícias verdadeiras.....</i>	69
7.4	RESULTADOS COM A TÉCNICA TF-IDF.....	70
7.5	RESULTADOS COM A TÉCNICA BAG OF WORDS.....	76
7.6	SALVANDO O APRENDIZADO DOS 3 MELHORES RESULTADOS EM ARQUIVO.PKL.....	81
7.7	FERRAMENTAS USADAS PARA CONSTRUÇÃO DO SITE.....	81
7.8	BANCO DE DADOS E MODELAGEM.....	81
7.8.1	<i>Modelo conceitual:.....</i>	81
7.8.2	<i>Modelo lógico.....</i>	82
7.8.3	<i>Dicionário de dados:.....</i>	82
7.9	CONSTRUÇÃO DO SITE.....	82
8	CONSIDERAÇÕES FINAIS.....	105

REFERÊNCIAS.....	107
-------------------------	------------

Lista de Figuras

Figura 1 - Reflexos das Fake News em cada geração.....	21
Figura 2 - Fontes de checagem de Fake News de acordo com cada geração.....	22
.....	
Figura 3 - Algoritmo SVM.....	31
Figura 4 - Camadas ocultas de um MLP.....	34
Figura 5 - Diagrama de Caso de Uso.....	42
Figura 6 - Diagrama de Classes.....	47
Figura 7 - Diagrama de Implementação.....	48
Figura 8 - Diagrama de Caso de Uso - Detecção da veracidade.....	49
Figura 9 - Diagrama de Caso de Uso - Consultar notícias já verificadas....	49
Figura 10 - Diagrama de Caso de Uso - Alterar Usuário.....	50
Figura 11 - Diagrama de Caso de Uso - Cadastrar usuário.....	50
Figura 12 - Diagrama de Caso de Uso - Deletar usuário.....	50
Figura 13 - Diagrama de Caso de Uso - Inserir notícia no bando de dados	51
Figura 14 - Diagrama de Caso de Uso - Login.....	51
Figura 15 - Diagrama de Sequências - Cadastro com sucesso.....	52
Figura 16 - Diagrama de Caso de Uso - Cadastro e-mail já cadastrado....	52
Figura 17 - Diagrama de Caso de Uso - Cadastrar erro na senha.....	53
Figura 18 - Diagrama de Caso de Uso - Login.....	53
Figura 19 - Diagrama de Caso de Uso - Login não encontrado.....	54
Figura 20 - Diagrama de Caso de Uso - Deletar.....	54
Figura 21 - Diagrama de Caso de Uso - Alterar.....	55
Figura 22 - Diagrama de Caso de Uso - Visualizar notícias cadastradas....	55
Figura 23 - Diagrama de Caso de Uso - Notícia falsa.....	56
Figura 24 - Diagrama de Caso de Uso - Notícia verdadeira.....	56
Figura 25 - Diagrama de Caso de Uso - Quantia de palavras não OK.....	57
Figura 26 - Diagrama de Caso de Uso - Língua Portuguesa não OK.....	57
Figura 27 - Diagrama de Atividades – Geração do Algoritmo com TF-IDF.	59

Figura 28 - Diagrama de Atividades – Geração do Algoritmo com Bag of Words.....	60
Figura 29 - Diagrama de Atividades – Atividade da detecção e gravação no bando de dados.....	61
Figura 30 - Diagrama de Estados – Atividade da detecção e gravação no bando de dados.....	62
Figura 31 - Dataset.....	67
Figura 32 - Gráfico com as 30 palavras com maior frequência:.....	68
Figura 33 - Nuvem de palavras do dataset.....	69
Figura 34 - Gráfico das 30 palavras mais frequentes falsas.....	69
Figura 35 - Nuvem de palavras das notícias falsas.....	70
Figura 36 - Gráfico das 30 palavras mais frequentes verdadeiras.....	70
Figura 37 - Nuvem de palavras das notícias verdadeiras.....	71
Figura 38 - Resultado da regressão logística com a matriz de confusão....	72
Figura 39 - Resultado da SVM com a matriz de confusão.....	73
Figura 40 - Resultado Naive Bayes MultinomialNB com a matriz de confusão.....	74
Figura 41 - Resultado Naive Bayes BernoulliNB com a matriz de confusão:.....	75
Figura 42 - Resultado MLPClassifier com a matriz de confusão:.....	76
Figura 43 - Resultado da regressão logística com a matriz de confusão:...	77
Figura 44 - Resultado da SVM com a matriz de confusão.....	78
Figura 45 - Resultado Naive Bayes MultinomialNB com a matriz de confusão:.....	79
Figura 46 - Resultado Naive Bayes BernoulliNB com a matriz de confusão.....	80
Figura 47 - Resultado MLPClassifier com a matriz de confusão:.....	81
Figura 48 - Modelo conceitual do banco de dados.....	82
Figura 49 - Modelo lógico do banco de dados.....	83
Figura 50 - Tela de cadastro e login.....	84

Figura 51 - Tela que auxilia a digitação da senha (quando está faltando algum requisito).....	85
Figura 52 - Tela que auxilia a digitação da senha (quando todos requisitos estão corretos).....	85
Figura 53 - Tela de cadastro quando falta preencher algum campo para o cadastro.....	86
Figura 54 - Tela de cadastro quando e-mail já foi cadastrado.....	86
Figura 55 - Tela cadastro quando as senhas digitadas são diferentes.....	87
Figura 56 - Tela cadastro quando a senha não atende os requisitos mínimos.....	87
Figura 57 - Tela cadastro quando a conta é registrada.....	88
Figura 58 - Tela de login quando a sua senha ou e-mail não são encontrados no banco de dados.....	89
Figura 59 - Página inicial.....	90
Figura 60 - Tela de análise quando a notícia não está língua portuguesa. .	91
Figura 61 - Tela de análise quando a notícia tem menos de 100 palavras. .	91
Figura 62 - Tela de análise quando a notícia tem o resultado de verdadeira:.....	92
Figura 63 - Tela de análise quando a notícia tem o resultado de falsa.....	92
Figura 64 - Tela do site com e-mail não compatível com a sessão e-mail criada.....	93
Figura 65 - Tela do site com e-mail alterado.....	94
Figura 66 - Tela do site com e-mail não compatível com a sessão e-mail criada.....	95
Figura 67 - Tela do site com senha não compatível com a sessão e-mail criada.....	96
Figura 68 - Tela do site com senha não a atendo aos requisitos mínimos. .	97
Figura 69 - Tela do site com senha alterada com sucesso.....	98
Figura 70 - Tela do site com e-mail não compatível com a sessão e-mail criada.....	99

Figura 71 - Tela do site com senha não compatível com a sessão senha criada.....	99
Figura 72 - Tela do site com conta deletada.....	100
Figura 73 - Tela histórico inicial.....	101
Figura 74 - Tela histórico com a notícia completa.....	101
Figura 75 - Tela esqueceu senha com e-mail não encontrado.....	102
Figura 76 - Mensagem Enviada para o e-mail.....	103
Figura 77 - Tela mudar a senha com senhas digitas diferentes.....	103
Figura 78 - Tela mudar senha com senha não atendendo requisitos mínimos.....	104
Figura 79 - Tela mudar senha alterada com sucesso.....	104
Figura 80 - Tela mudar senha sem sessão.....	105

1 Introdução

Com a facilidade de disseminação de notícias que se tem com as redes sociais e aplicativos de comunicação, as notícias falsas criaram grande força para influenciar pessoas e gerando um grande desserviço social. (VEJA, 2020) 62% dos brasileiros não sabem identificar uma *Fake News* e mesmo assim temos uma porcentagem melhor que muitas países da América Latina.

(O PERIGO... 2021) Com a popularização e fácil acesso aos meios de comunicação e ficando cada vez mais fácil o compartilhamento de informações mentirosas as consequências são gigantescas. No Brasil, em 2014, o compartilhamento de uma *Fake News* causou uma grande tragédia. No caso, uma mulher foi linchada até a morte por moradores da cidade de Guarujá, em São Paulo. Fabiane Maria de Jesus tinha 33 anos, era dona de casa, casada, mãe de duas crianças, e foi associada a uma sequestradora de crianças, com um retrato falado feito dois anos antes.

(O PERIGO... 2021) Mais um caso famoso das consequências das *Fake News*, é do movimento anti-vacinação. Em que pessoas contrárias às vacinas, propagam notícias falsas, falando que a composição química das vacinas causa problemas de saúde. A consequência desse foi o alto crescimento de sarampo no Brasil em 2018.

Diante do mal que as notícias falsas causam na sociedade, esse trabalho tem como objetivo identificar *Fake News*, via aprendizado de máquina com o algoritmo de regressão logística, utilizando a técnica *TF-IDF* e assim criar uma aplicação com a capacidade de prever a veracidade das notícias para que aqueles que façam sua utilização, se sintam mais confortável na recepção e compartilhamento de notícias.

Com as pesquisas bibliográficas, estudo de *Python*, *Flask*, *SQL*, *Machine Learning*, mineração de dados, algoritmos preditivos, foi possível a criação da aplicação apresentada que recebe notícia de um usuário e fará a predição da notícia, graças ao *Machine Learning* e o algoritmo preditivo de regressão logística.

1.1 Objetivo

O objetivo deste trabalho é criarmos uma página que possa identificar a veracidade de notícias sendo elas falsas ou verdadeiras.

1.2 Objetivo Geral

O objetivo deste Trabalho de Graduação é desenvolver um site através de *machine learning* utilizando algoritmos de predição de redes neurais, voltado para todo e qualquer tipo de usuário que consome notícias seja pela internet, mensagens, site, blogs etc., a fim de mitigar os possíveis efeitos que uma *Fake News* possa causar na vida das pessoas. Além do mais, este site servirá de apoio para pesquisadores, fontes de pesquisas, portadores de notícias e claro todos nós em nosso dia a dia.

1.3 Justificativa e motivações

Diante do cenário que estamos acompanhando, o termo *Fake News* vem ganhando muita popularidade por conta do quão prejudicial esse tema possa ser e seus impactos na vida das pessoas, a partir disso surgiu-se a ideia de que fosse construído uma ferramenta que pudesse mitigar tanto a propagação de notícias falsas quanto as próprias situações de enganações onde muitas pessoas já passaram ou poderão passar. Mediante a isso criamos este projeto para contribuir da melhor maneira para a sociedade.

1.4 Organização do trabalho

Este trabalho é composto por 9 capítulos, que estão distribuídos em introdução onde abordaremos os objetivos e justificativas deste trabalho, após este capítulo temos a proposta de desenvolvimento de projeto trazendo as informações coletadas a respeito das *Fake News*, o pré-processamento e processamento, os algoritmos, técnicas e linguagens utilizadas na construção do projeto.

No terceiro capítulo trazemos todos os requisitos presentes no sistema do software, abrangendo aqueles que são funcionais e não-funcionais, além de apresentarmos toda a parte de modelagem funcional do projeto e os protótipos. No capítulo seguinte temos nosso caso de uso e o quinto capítulo que complementa o anterior com os demais diagramas e toda a arquitetura do sistema.

Após este capítulo temos os testes que envolve os planos e os roteiros, no próximo capítulo os resultados dos algoritmos utilizados e não utilizados e por fim as considerações finais seguida das referências que nos baseamos para a criação deste projeto.

2 Proposta De Desenvolvimento Do Projeto

Neste capítulo será apresentado as ferramentas utilizadas para desenvolvimento do projeto, explorando desde o pré-processamento e o próprio processamento realizado através do aprendizado de máquina, algoritmos preditivos de redes neurais e as linguagens de programação utilizadas ao longo do projeto.

2.1 O que é Fake News

Segundo Galhardi, Freire, Minayo e Fagundes (2020), *Fake News* tem denominação a produção e propagação em grande escala a de notícias falsas, feitas para distorcer fatos de forma intencional, enganar, desinformar, induzir a erros, manipular a opinião pública, depreciar ou engrandecer uma instituição ou uma pessoa, diante de um assunto específico, para obter vantagens econômicas e políticas.

As mentiras com alta disseminação não são algo novo. Muito antes da internet, histórias como “Elvis não morreu” ou de o homem não ter pisado na lua já circulavam na sociedade e uma parcela grande da população tomava essas notícias como verdade (ALVES; MACIEL, 2020).

As desinformações têm um objetivo claramente político. Um bom exemplo disso são as falsas estações de rádio alemãs, com transmissões no Reino Unido durante a Segunda Guerra Mundial, em que o interlocutor inglês se passava de alemão e transmitia comentários contrários contra o líder nazista Adolf Hitler (ALVES; MACIEL, 2020).

Em 2016 o termo *Fake News* obteve uma crescente popularização, durante a cobertura das eleições americanas. O termo foi muito usado pela mídia pelos candidatos à presidência, visando desqualificar informações que favorecessem a si próprios (GALHARDI; FREIRE; MINAYO; FAGUNDES, 2020).

2.1.1 A presença de Fake News em redes sociais

Diante do cenário que temos hoje, notícias falsas estão cada vez mais presente em nossas vidas pois ela se espalha justamente através de um local que podemos dizer que a humanidade se tornou refém. A desinformação toma e chega a lugares onde jamais possa se imaginar, seria ingenuidade descrevermos a internet atualmente somente como um espaço de comunicação autônomo, nela qualquer transmissão pode gerar uma reação positiva ou negativa sendo colocada em diferentes contextos por diferentes pessoas, gerando assim uma desinformação que podendo ser compartilhada formando possivelmente uma bola de neve que podemos chamar de *Fake News*.

Segundo uma pesquisa de uma empresa internacional especializada em pesquisas sobre internet (YouGov), entre os meses de junho e julho, pessoas de sete regiões diferentes do mundo foram entrevistadas, sendo os seguintes países: Estados Unidos, Brasil, Reino Unido, Alemanha, Nigéria, Índia e Japão. Foram consideradas também a diferença de geração entre elas integrando a Geração Z (18-25 anos), *Millenials* (26-41 anos), Geração X (42-57 anos), Boomers (58-67 anos) e Geração “Silenciosa” (68+ anos). (MANNARA, 2022)

De acordo com a pesquisa os países mais propensos a terem acesso a informações falsas ou enganosas são EUA, Reino Unido, Brasil e Nigéria. No Brasil especificamente 44% afirmam ter contato com *Fake News* diariamente, enquanto 27% e 13% afirmam ser impactados semanalmente e mensalmente, respectivamente. Além disso, em todos os países combinados 62% das pessoas analisadas acreditam receber informações enganosas semanalmente. (MANNARA, 2022)

Algo que trouxe muita preocupação entre os entrevistados é que de 4 pessoas 10 estão muito preocupadas com os efeitos que informações enganosas possam causar na vida das pessoas, a preocupação maior é voltada para os jovens da geração Z e *Millenials*, pois são crianças/jovens que estão frequentemente em contato com a internet, além de que são conhecidas como geração “*fast-food*” por muitos, por querer tudo muito rápido na hora que quer,

fazendo com que utilizem a primeira informação que encontram pela frente. (MANNARA, 2022). Figura 1 Reflexos das Fake News em cada geração.

Figura 1 - Reflexos das Fake News em cada geração



Fonte: YouGov

Ainda que a geração Z e *Millennials* sejam consideradas a geração “*fast-food*”, são as gerações que mais checam com mais frequência a veracidade das informações consumidas comparadas as outras gerações de idades mais avançadas. De acordo com a pesquisa as principais checagens são feitas em sites de pesquisas (44%), como Google ou Bing, sendo que nesse quesito as gerações mais jovens sentem mais confiantes em realizar a identificação dos conteúdos falsos do que as pessoas mais velhas. (MANNARA, 2022)

Os usuários entrevistados também alegam a checagem através de aplicativos como WhatsApp (39%), *Facebook* (36%), *YouTube* (34%) e sites de notícias nacionais (33%), entre os veículos que as pessoas mais confiam, ou seja, aqueles que menos verificam informações recebidas estão *Twitter* (22%), TikTok (16%) e Snapchat (12%). (MANNARA, 2022)

Entre todas as gerações, a forma mais ampla de checar a fonte das notícias é feita por meios de comunicações com fortes reputações (49%), ou seja, amplamente conhecida como jornais e revistas, após isso o que preocupa grande parte dos usuários são questões que envolvem as datas de postagem das notícias (46%) e por fim buscam saber mais sobre a informação e quem a publicou. Figura 2 Fontes de checagem de Fake News de acordo com cada geração.

Figura 2 - Fontes de checagem de Fake News de acordo com cada geração

Source of information	All	Gen Z	Millennials	Gen X	Boomers	Silent
A search engine, such as Google or Bing	44%	51%	49%	45%	39%	29%
A messaging app, such as WhatsApp, Telegram or Signal	39%	48%	46%	40%	28%	19%
Facebook	36%	38%	42%	37%	32%	22%
A video platform, such as YouTube or Vimeo	34%	44%	41%	35%	26%	15%
National news websites	33%	32%	38%	33%	31%	22%
Local news websites (for your city or county)	31%	32%	37%	31%	30%	20%
A news app	30%	31%	34%	32%	28%	21%
Instagram	29%	44%	39%	27%	14%	7%
Online blogs	25%	33%	31%	24%	19%	11%
Twitter	22%	30%	28%	22%	13%	6%
TikTok	16%	28%	22%	15%	7%	3%
Snapchat	12%	20%	16%	9%	5%	2%
Other	13%	22%	20%	14%	5%	3%

Fonte: YouGov

2.1.2 Consequências das Fake News

A desinformação e a propagação de informações enganosas podem causar sérios problemas na sociedade destruindo reputações, privacidade, podendo gerar violência para as pessoas envolvidas em situações como essas, além de trazer discriminação e hostilidade com diversos grupos distintos em nossa sociedade. (CARRIÇO; PIRES; TERRA; BASILIO, 2019)

Outras consequências que podem ocorrer a partir do compartilhamento de informações são aquelas que prejudicam o próprio indivíduo, apresentando riscos a seus direitos de liberdade e expressão em casos que a autoridade pública possa denigrir ou tentar impedir a disseminação de notícias falsas. (CARRIÇO; PIRES; TERRA; BASILIO, 2019)

Casos recentes que tivemos como em março de 2020 em que durante a pandemia, devido ao compartilhamento de informações equivocadas 27 pessoas morreram intoxicadas no Irã após ingerirem álcool adulterado como forma de tratamento da COVID-19, após acreditarem num boato falso de que bebidas alcoólicas ajudariam a combater o vírus. (AFP, 2020)

Outro caso que testemunhamos foi em 2018 quando cresceu-se os casos de Sarampo no Brasil, resultando numa intensa campanha de conscientização realizada pelo Ministério da Saúde, com o objetivo de combater a disseminação de Fake News que surgiram na época, foram lançados propagandas e informativos

de combate às falsas informações sobre a vacinação em diversos veículos de informações e redes sociais, porém mesmo assim muitas pessoas não confiaram o suficiente nas informações corretas resultando na abertura de precedentes para a não vacinação de diversas pessoas que em situações epidêmicas são um grande perigo para a sociedade. (TJPR, 2020)

2.1.3 Como se prevenir das Fake News

Sabemos que é muito mais fácil e de certa forma muitas vezes mais cômodo, utilizarmos informações que podem ser ou não confiáveis, pelo fato de clicarmos no primeiro site em que vemos, porém por tudo o que vemos e enfrentamos em nosso dia a dia, não podemos mais dar margem para estas ações que podem se tornar grandes erros, para isso algumas formas de evitar que novos consumos de falsas informações ocorram podemos seguir alguns passos como: (TJPR, 2020)

- Confirmar diferentes fontes de uma determinada notícia, sempre bom duvidar de uma notícia principalmente de sites que não são muito conhecidos, levando até mesmo a checagem dos autores de determinada notícia;
- Conferir as datas da publicação, acontecem muitos casos em que a notícia pode tratar do assunto a procura, mas possa ter ocorrido de formas diferentes ou enganosas em outros períodos.

2.2 Mineração de Dados

Mineração de dados trata-se de extrair ou trabalhar conhecimento de grandes volumes de dados. O termo mineração de dados também é considerado sinônimo de *Knowledge Discovery in Databases* (KDD) ou descoberta de conhecimento em banco de dados. E o KDD é um processo que consiste nas seguintes etapas:

1. Limpeza de dados: etapa em que são retirados dados inconsistentes;
2. Integração dos dados: etapa onde em que fontes diferentes de dados são combinadas gerando uma única base de dados;

3. Seleção: etapa onde que são selecionados os atributos de interesse ao usuário. Por exemplo, o usuário pode decidir quais informações são relevantes ou não que defina se um cliente é um bom comprador ou não;

4. Transformação dos dados: etapa em que os dados são transformados em um formato apropriado para aplicação de algoritmos de mineração;

5. Mineração: etapa de grande importância para o projeto, consistindo na aplicação de técnicas inteligentes com o objetivo de extrair os padrões que tenham interesse;

6. Avaliação ou Pós-processamento: etapa que é identificado os padrões interessantes de acordo com algum critério do usuário.

7. Visualização dos Resultados: essa última etapa é onde se utiliza técnicas de representação de conhecimento com a intenção de apresentar ao usuário o que foi obtido com a mineração, (AMO, 2004).

2.2.1 Tarefas e Técnicas de Mineração de Dados

Segundo Amo (2004), é importante saber a diferença de técnica de mineração. A tarefa é a especificação do dado que estamos querendo buscar, buscando regularidades ou padrões que temos interesse em encontrar, ou que tipo de padrões que podem abrir um sinal de alerta (por exemplo, um gasto exagerado de um cliente de cartão de crédito).

A técnica de mineração tem a função de especificar de métodos que possibilite descobrir os padrões que nos interessam. As principais técnicas utilizadas em mineração de dados, se tem técnicas estatísticas, técnicas de aprendizado de máquina.

2.2.2 Processamento de Linguagem Natural

Processamento de linguagem natural (PLN) é uma vertente da inteligência artificial que ajuda computadores a entender, interpretar e manipular a linguagem humana. O PLN é a mescla de diversas disciplinas que envolvem ciência da

computação e a linguística computacional, com o objetivo de evitar e preencher os espaços falhos que ocorrem entre a comunicação humana e o entendimento dos computadores a partir da linguística computacional. (PROCESSAMENTO...)

O PLN vem em um grande crescente neste século ainda mais que temas que envolvem Inteligência artificial, *big data*, comunicação homem-máquina, aprimoramento de algoritmos, fatores que incentivam o crescimento de processamento em níveis de predição. (PROCESSAMENTO...)

Uma das grandes importâncias do PLN para todos nós é pelo fato de que este processamento atua como o meio campo entre homem e máquina escalando outras tarefas à linguagem. Em comparação aos seres humanos, o trabalho que as máquinas fazem são de extrema ajuda, pois elas podem analisar mais dados baseados do que nós humanos sem fadiga, de maneira constante e imparcial.

Em termos gerais, as tarefas do PLN segmentam a linguagem em partes menores e essenciais, tentando entender como elas se relacionam, explorando e estruturando cada parte com o objetivo de que traga significado, entendimento e funcionamento para aquilo que está sendo tratado. (PROCESSAMENTO...)

2.2.3 Feature Extraction

Para que possa ser possível utilizar um modelo estatístico ou de *deep learning* em NLP, é necessário *features*: que são informações mensuráveis acerca de alguma ocorrência, que é uma forma estruturada de armazenar dados. Mas textos é um tipo de dado não estruturado (não organizado de uma maneira pré-definida), sendo assim, é difícil para o computador entendê-los e analisá-los. Por isso, é necessário utilizar *feature extraction*, então transformando o texto em uma informação numérica de modo que seja possível utilizá-lo para alimentar um modelo, (FONSECA, 2020).

2.2.4 Stopwords

Normalmente as *stopwords* são consideradas palavras irrelevantes seja numa frase ou num texto em que durante o pré-processamento a PLN atua

identificando estas palavras e eliminando àquelas que não são de tanta utilização em questão de informações e relevância semânticas, ao longo do texto. Entre as *stopwords* encontradas, são elas identificadas através de preposições, pronomes ou conjunções. (MARCEL, 2009)

2.2.5 Tokenização

Outra etapa muito importante e frequente que ocorre durante o pré-processamento de PLN é a “Tokenização” que tem a função de pegar as palavras que são encontradas e distribuídas no texto, armazenando-as em uma lista. (HEISE, 2020)

Em nosso algoritmo utilizamos a biblioteca NLTK (*Natural Language Toolkit*) disponível na linguagem *Python* para a realização de tarefas em PNL a partir de análises e processamentos textuais. (VITORIA, 2020)

2.2.6 Stemming

A técnica de stemização consiste em reduzir palavras relacionados em um texto a uma forma mínima para que possam ser combinadas sob uma única forma de apresentação chamada, *stem*. Com esta técnica aumenta se a possibilidade de termos resultados mais precisos focando na captura da essência das palavras e de suas diversas variações, podendo então melhorar a qualidade dos resultados e a simplificação das informações. (COELHO, 2007)

Durante o processo de stemização dois tipos de erros podem ocorrer com frequência o primeiro seria o *overstemming* no qual acontece a remoção de mais letras/palavras indevidamente, permitindo que o sentido ou diferentes palavras apontem para o mesmo *stem*; o segundo erro que pode ocorrer é o *understemming* que é o contrário do primeiro caso no qual ocorre a não retirada de letras/palavras da maneira correta as deixando para trás e conseqüentemente criando *stems* diferentes, porém com o mesmo significado. (ALVARES, 2014)

2.2.7 TF-IDF

O algoritmo *TF-IDF* (*term frequency-inverse document frequency*) utiliza técnicas estatísticas durante o processo de mineração de texto, recuperação de informações e o processamento de linguagem natural, com sua principal função de definir o grau de importância das palavras presentes em um texto, de acordo com a quantidade de vezes de sua aparição, seja em textos estruturados ou semi-estruturado (STEIN; SILVA, 2016)

Sua equação é representada da seguinte maneira: TF representado por $TF(i, j)$, que é o número de vezes em que o termo i aparece no documento j . Enquanto o IDF representado por $IDF(i, k)$, é o logaritmo do total de documentos (k) dividido pelo número de documentos que contém o termo i . (KIDO; MORIGUCHI; JUNIOR, 2014)

$$TF-IDF(i, j, k) = TF(i, j) \times IDF(j, k)$$

Essa técnica tem como objetivo de filtrar um texto, classificando as palavras com valores que representam um grau de afinidade correlato ao texto. Termos que são mais comuns localizados ao longo do texto, tendem a ter um maior TF-IDF, ou seja, não possuem uma grande importância para o texto por sua presença que ocorrem mais vezes que outras palavras, sendo o caso de pronomes, artigos e preposições. (KIDO, 2014; MORIGUCHI, 2014; JUNIOR, 2014)

2.2.8 Bag of words

Bag of Words é uma das formas de representar o texto de acordo com a ocorrência das palavras nele. Traduzindo para o português, o “saco de palavras”, ele recebe esse nome por não levar em conta a ordem ou a estrutura das palavras no texto, leva em conta apenas se a palavra aparece ou a frequência ela aparece no texto. (FONSECA, 2020).

Um exemplo, se uma palavra aparece muito num texto, ela se torna mais importante para a máquina. Assim *Bag of Words* é uma ótima alternativa para

determinar as palavras que possuem o maior significado de um texto com base no número de vezes que ela se repete. (FONSECA, 2020).

2.3 Inteligência artificial

A inteligência artificial tem como método ou procedimento que é realizado por uma máquina e em uma tomada de decisão, que são características de hábitos ou atitudes produzidas pelo um ser portador de inteligência. E nisso envolve conceitos que são: otimização, reconhecimento de padrões, automatização, robótica etc. (GUIMARÃES, 2016)

2.3.1 Machine Learning

O *Machine Learning*, mais conhecido como Aprendizado de Máquina tem como objetivo a criação de programas que tenham bons desempenhos por exemplos fornecidos por uma grande quantidade de dados. O aprendizado de máquina é orientado a dados, ele aprende automaticamente a partir de grandes volumes de dados. Os algoritmos conseguem gerar hipóteses a partir dos exemplos dos dados fornecidos. (LUDERMIR, 2021).

Existe, 3 tipos de modelos de aprendizado de máquina que são: Supervisionado, Não Supervisionado e por reforço. (LUDERMIR, 2021)

2.3.2 Supervisionado

No modelo Supervisionado, para cada dado apresentado ao algoritmo de aprendizado é necessário mostrar a resposta desejada (um rótulo informando a que classe o exemplo pertence, no caso de um problema de classificação de imagens, por exemplo, como distinguir imagens de gatos e de cachorros). (LUDERMIR, 2021)

2.3.3 Não supervisionado

O Não Supervisionado, os dados são entregues ao algoritmo sem rótulos. O algoritmo fará o agrupamento dos dados pelas similaridades dos seus atributos. O

algoritmo analisa os dados entregues e tenta determinar o agrupamento deles de alguma maneira, formando agrupamentos ou clusters. (LUDERMIR, 2021).

2.3.4 Por reforço

O modelo por Reforço, o algoritmo não recebe a resposta correta, mas recebe um sinal de reforço. Ele tenta fazer uma hipótese baseado nos dados e determina se essa hipótese é correta ou não. Ele é bastante utilizado em jogos e robótica.

2.3.5 Deep Learning

O *Deep Learning* que é uma das subáreas de inteligência artificial tem como base uma tecnologia que representa redes neurais, para trazer semelhanças ao funcionamento do cérebro humano para compreender, raciocinar, analisar e concluir dados geridos dinamicamente a partir de diferentes tipos de dados que atuam em camadas hierárquicas no processamento de informações permitindo um alto nível de abstração para maiores análises e mais complexas sobre cada dado. (TOTVS, 2020)

Um dos principais motivos do uso do *Deep Learning* é o fato de que suas redes neurais são utilizadas para a revelação de *insights* e descobertas que poderiam até existir, mas não eram visualizados de maneira clara e efetiva de como este aprendizado possibilita. Com modelos de machine learning mais robustos, é possível que as empresas possam analisar dados grandes e complexos, ajudando a melhorar detecção de fraudes, melhorar seguranças cibernéticas, prever resultados, trazendo cada vez mais eficiência e economia de tempo. (ORACLE, 2022)

2.4 Algoritmos de Predição

Nesta etapa abordaremos os diferentes algoritmos de predição que estão presentes em nosso projeto. Existem diversos modelos de predição alguns deles

como SVM, *Naive Bayes*, *MLPClassifier* e um mais conhecido pela maioria que é o de Regressão Logística serão abordados.

A escolha do algoritmo ideal para um projeto irá depender de diferentes fatores até levar a sua escolha, onde são levados em consideração qualidade do algoritmo, necessidades do projeto e usuário, eficiência, tipos de dados que são apresentados e do que será resolvido, além de outros fatores que possam se encaixar melhor. (MEDIUM, 2019)

Como o próprio nome diz, os algoritmos envolvidos têm como objetivo de prever possíveis resultados através de cálculos estatísticos e probabilísticos, sendo usados em diversas áreas com diferentes propósitos e objetivos.

2.4.1 Regressão Logística

A Regressão logística, é uma técnica que calcula a probabilidade de obter um resultado, sendo assim, ela tem a capacidade de obter a probabilidade de um evento ocorrer. A Regressão Logística é muito utilizada na área da saúde, além disso o fato de possuir uma larga eficiência, faz com que seja viabilizada em mais áreas diversas, que vai de ciências médicas até avaliação de créditos. (GONZALEZ, 2018)

Como abordado anteriormente, a regressão logística devido a sua eficiência inclui todos os campos da ciência média e sociais, sendo encontradas na política, através de pesquisas com o objetivo de prever possíveis resultados que dependem de variáveis que envolvem sexo, idade, local de residência, condições sociais e padrões de votações anteriores no qual a partir dessas variáveis são utilizados métodos de regressão logística por meio de probabilidades para a previsão dos resultados. (TIBCO, 2022)

Além de ser usados em áreas como marketing, testes de produtos, setores financeiros, comércios eletrônicos, diversas áreas com um objetivo principal em comum, prever resultados para uma melhor satisfação dos usuários buscando sanar seus problemas e dificuldades. (TIBCO, 2022)

Podemos distinguir a regressão logística em 3 tipos básicos, sendo elas:

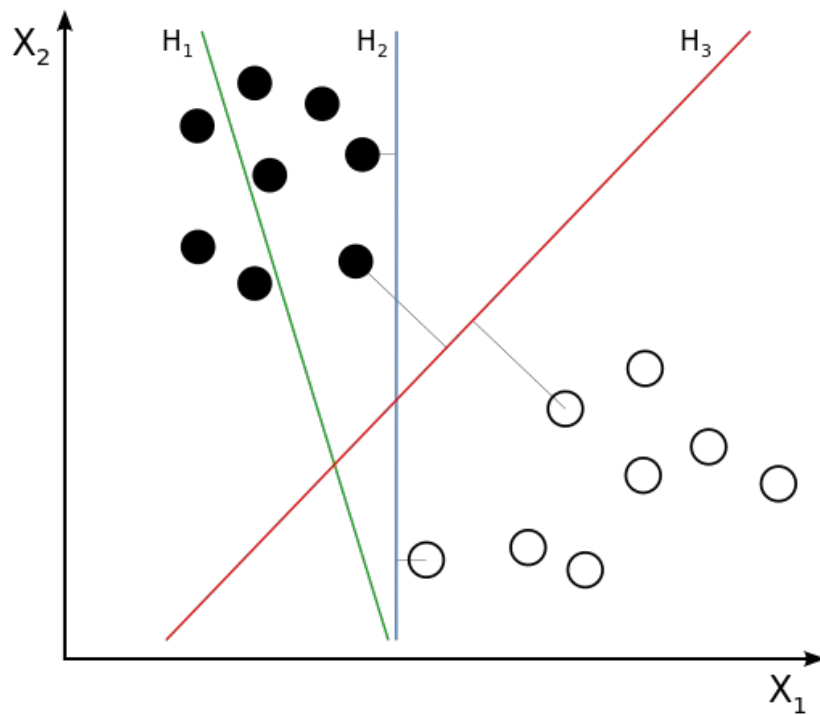
- **Binária** onde há apenas duas categorias para a resposta de uma variável;
- **Multinomial** no qual podem ser encontrados mais de 3 respostas distintas que não possuem nenhuma ordem para uma determinada variável;
- **Ordinal** que se assemelha a Multinomial possuindo 3 ou mais respostas em uma variável, porém de forma ordenada.

2.4.2 SVM

O SVM (*Support Vector Machine*) é uma técnica de classificação baseada em aprendizado de máquina na qual é utilizado o aprendizado supervisionado que pode ser usado para classificação e regressão. (ADDAN, 2019)

O SVM é um algoritmo que busca uma linha de separação de 2 classes distintas analisando dois pontos um de cada grupo distinto mais próximos da outra classe, ou seja, o SVM escolhe o hiperplano (reta) em maiores dimensões entre dois grupos que se distancia de cada um. Após descoberta desta reta, o algoritmo consegue predizer qual a classe um novo dado pertence a partir de qual lado da reta ele está. (COUTINHO, 2019). Figura 3: Algoritmo SVM

Figura 3 - Algoritmo SVM



Fonte: Wikimedia Commons

2.4.3 Naive Bayes

Este algoritmo se preocupa em problemas de classificação através de um embasamento estatístico de predições onde o algoritmo define uma tabela de probabilidades, constando a frequência dos preditores em relação as variáveis de saída por meio de *machine learning*. Durante seu pré-processamento os cálculos para a realização da previsão são independentes entre si, assumindo também que as variáveis *features* são iguais e muito importantes para o resultado final. (SACRAMENTO, 2021)

O classificador é tido como *Naive* (ingênuo) pois assume que a informação de um determinado evento, não serve de informação para outro evento inserido no mesmo contexto. As principais razões da utilização desse classificador ser utilizado é por conta da sua facilidade e rapidez para implementar e apresentar, trazendo assim uma boa eficácia, relativamente. (QUIXADÁ, 2019)

Importante se destacar que existem 3 tipos de modelo *Naive Bayes*, sendo eles o Gaussian que é usado na classificação assumindo uma distribuição normal, Multinomial usado para contagem e o Bernoulli que faz o uso de incidência de eventos.

$$P(y|x_1, \dots, x_n) = P(y) \prod_{i=1}^n P(x_i|y) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)}$$

2.4.3.1 Bernoulli Naive Bayes

O modelo Bernoulli faz o uso de incidência de eventos, sua base de treinamento é composta por um vetor binário que contém a palavra e a sua incidência dentro do conjunto de treinamento, este vetor binário armazena todas as palavras encontradas durante o treinamento. (PIVETTA, 2013)

Além do mais este algoritmo se baseia numa base de acordo com a Distribuição Multivariada de Bernoulli, que são compostas por diversas *features*, as quais são valores binários. (TAMAI, 2019)

$$P(x_i|y) = P(i|y)^{x_i} (1 - P(i|y))^{1-x_i}$$

2.4.3.2 Multinomial Naive Bayes

O modelo multinomial que utiliza a função MultinomialNB implementa o algoritmo *Naive Bayes* para dados multinomialmente distribuídos, sendo a segunda variante clássica deste algoritmo usado para a classificação de textos no qual é representado tipicamente como contagem de vetores de palavras presentes ao longo do texto, possuindo um bom funcionamento assim como o TF-IDF. (SCIKIT-LEARN DEVELOPERS, 2022)

A distribuição do multinomial é feita com a distribuição que é parametrizada por vetores $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$ para cada classe y onde n é o número de *features* (na classificação de texto e tamanho do vocabulário) e θ_{yi} é a probabilidade de $P(x_i|y)$ da *feature* i aparecendo numa amostra da classe y . (SCIKIT-LEARN DEVELOPERS, 2022)

O parâmetro θ_y é estimado por uma versão máxima de probabilidade, por exemplo uma contagem de frequência relativa. (SCIKIT-LEARN DEVELOPERS, 2022)

$$\hat{\theta}_{yi} = \frac{N_{yi} + a}{N_y + an}$$

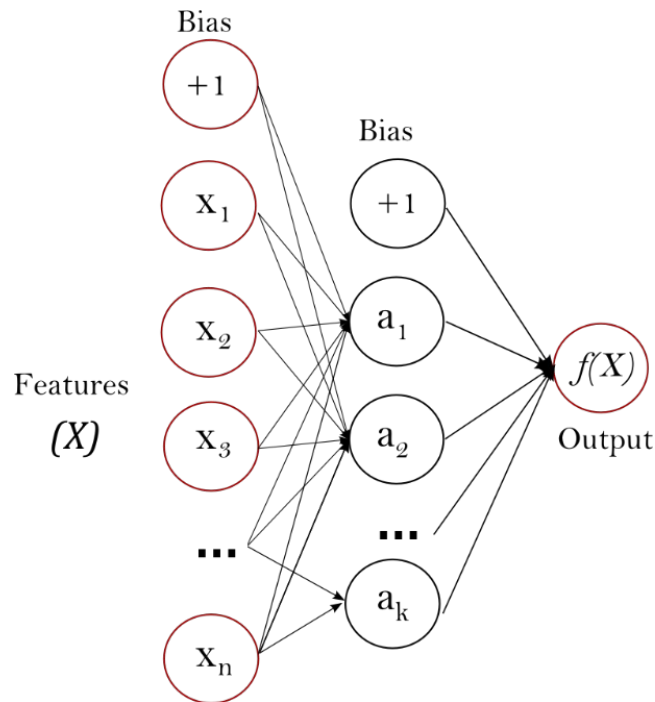
2.4.4 MLPClassifier

O *Multi-layer Perceptron* (MLP) é um algoritmo de aprendizado supervisionado que aprende a função $f(\cdot): R^m \rightarrow R^o$ através do seu treinamento dentro do *dataset*, onde m é número de dimensões para entrada e o para saída. (SCIKIT-LEARN DEVELOPERS, 2022)

Dado um conjunto de *features* $X = x_1, x_2, \dots, x_m$ e um objetivo y , o algoritmo pode aprender e captar uma função aproximada não linear tanto para classificação quanto para regressão. (SCIKIT-LEARN DEVELOPERS, 2022)

Acaba sendo diferente para casos de regressão logística, pelo fato de que entre a camada de entrada e de saída pode haver camadas não lineares que são chamadas de camadas ocultas como o exemplo a seguir. (SCIKIT-LEARN DEVELOPERS, 2022). Figura 4 como são as camadas

Figura 4 - Camadas ocultas de um MLP



Fonte: Scikit-learn

2.5 Recursos

Para o desenvolvimento deste projeto, nosso *BackEnd* foi feito utilizando a linguagem *Python*, integrando o framework Flask, utilizamos o banco de dados relacional MySQL para armazenarmos as informações dos usuários e seus históricos de pesquisas que podem ficar guardadas por conta do banco de dados. Em relação ao *FrontEnd*, utilizamos recursos como *JavaScript*, HTML, CSS e *Bootstrap* para fazer toda a interface de utilização do usuário.

2.5.1 Python

A linguagem de programação *Python* foi criada em 1990 por Guido van Rossum, tinha como foco original usuários como físicos e engenheiros. O *Python* foi concebido a partir de outra linguagem existente na época, chamada ABC. E hoje é utilizada por várias empresas grande de tecnologia, como Google e Microsoft. (BORGES, 2009)

Ela inclui diversas estruturas de alto nível (listas, tuplas, dicionários, data / hora, complexos e outras) e uma vastidão de módulos prontos para uso, além de *frameworks*. Uma linguagem que suporta programação modular e funcional, além da orientação a objetos. A linguagem é interpretada através de *bytecode* pela máquina virtual *Python*, tornando o código portátil. (BORGES, 2009)

Python é utilizado vários *softwares*, permitindo automatizar tarefas e adicionar novas funcionalidades, entre eles: PostgreSQL, Blender e GIMP. Também tem a possibilidade de integrar com outras linguagens. E apresenta muitas similaridades com outras linguagens dinâmicas, como Perl e Ruby. (BORGES, 2009)

2.5.2 Flask

Lançado em 2010 e desenvolvido por Armin Ronacher, o *Flask* é um *micro-framework* para pequenas aplicações com requisitos mais simples. Permite que um projeto possua apenas os recursos necessários para sua execução. Ele tem simplicidade porque possui apenas o necessário para o desenvolvimento de aplicações, tem uma arquitetura muito simples os projetos feitos em *Flask* normalmente são menores e mais leves em comparação a *frameworks* maiores. Andrade (2022)

2.5.3 JavaScript

Conhecido como linguagem ao lado do servidor o *JavaScript* chegou ao *Netscape Navigator 2.0* em setembro de 1995, obtendo um sucesso imediato e no ano seguinte foi introduzido o suporte da linguagem chamada JScript ao Internet Explorer 3.0. (MDN WEB DOCS, 2022)

O JavaScript é uma linguagem de programação utilizada principalmente para Scripts dinâmicos ao lado do cliente em páginas web, podendo ser utilizado ao lado do servidor com auxílio de um interpretador como o *NodeJS* por exemplo. Esta linguagem é muito utilizada principalmente em navegadores permitindo a manipulação dos conteúdos por meio do DOM, manipulação de dados com AJAX

e o IndexedDB, desenhos de gráficos fazendo com que os dispositivos interajam entre si, sendo executados no navegador através de APIs, entre outras. (MDN WEB DOCS, 2022)

2.5.4 Banco de dados Relacional

Um banco de dados do tipo relacional é aquele que armazena e fornece informações e dados que se relacionam entre si, baseado no modelo relacional. Representando uma maneira direta da amostragem dos dados em forma de tabela. Em um banco de dados Relacional, cada linha registrada possui um registro no qual possui um ID exclusivo que representa uma chave. Cada coluna da tabela contém os atributos dos dados e cada registro tem seu valor, o Banco de dados Relacional atua facilitando o relacionamento entre sim. (ORACLE, 2022)

Os bancos de dados relacionais são utilizados em diversas áreas/empresas por conta da sua eficiência com a procura e localização de dados independente da variedade e necessidades de informações. Eles podem ser usados para localizar inventários, processar transações, gerenciar grandes quantidades de informações de clientes, usuários e muito mais. (ORACLE, 2022)

Bancos de Dados relacionais existem desde os anos de 1970. E por conta de suas vantagens do modelo relacional, faz com que o modelo continue sendo aceito e integrado cada vez mais nas grandes empresas. (ORACLE, 2022)

2.5.5 MySQL

O MySQL é um sistema gerenciador de banco de dados relacional no qual é utilizado na maioria das aplicações de maneira gratuita com o intuito de gerir bases de dados. O sistema foi desenvolvido pela empresa MySQL AB e publicado originalmente em maio de 1995, após isso a empresa foi comprada pela Sun *Microsystem* e anos depois em 2010 foi integrado pela Oracle Corporation em uma transação da compra da Sun. (PISA, 2012)

O Serviço utiliza a linguagem do tipo SQL (*Structure Query Language* – Linguagem de Consulta Estruturada) sendo a linguagem mais popular no mundo

para inserção, acesso e gerenciamento de conteúdo armazenados num banco de dados. (PISA, 2012)

2.5.6 HTML

Criado em 1991, por Tim Berners-Lee, durante o CERN (*European Council for Nuclear Research*) na Suíça, o HTML inicialmente foi projetado para interligar instituições de pesquisa próximas, e compartilhar documentos com facilidade. Um ano depois foi liberado a biblioteca de desenvolvimento WWW (*World Wide Web*) que juntas proporcionaram o uso em escala mundial da WEB. (PACIEVITCH, 2019)

O HTML (Linguagem de Marcação de Hipertexto) é o bloco de construção mais básico da web usado para estruturar uma página web sem conteúdo. O “Hipertexto” refere-se aos *links* que conectam páginas da web entre si, seja em um único site ou entre outros, com isso pode-se ver a importância e o aspecto fundamental que os Links possuem na web. (MDN WEB DOCS, 2022)

2.5.7 CSS

O CSS (*Cascading Style Sheets*) é uma linguagem de estilo usada para descrever a apresentação de um código descrito em HTML ou em XML. Criado em 1995 por Håkon Wium Lie e Bert Bos, foi apresentada a proposta do CSS que logo foi apoiada pela W3C (World Wide Web Consortium). (ARTIGO... 2022)

No geral a ideia era utilizar o HTML somente para a estruturação de websites e a tarefa de apresentação ficaria sob responsabilidade do próprio CSS localizado em um arquivo separado .css ou no próprio HTML demarcado através de tags.

Seus conceitos de estilização na maioria das vezes não são seguidos totalmente, por conta de problemas de compatibilidades entre navegadores e também pela falta de conhecimento de alguns desenvolvedores. (ARTIGO... 2022)

2.5.8 Bootstrap

O *Bootstrap* é um *framework front-end* que fornece estruturas de CSS para a criação de sites, aplicações e páginas de dispositivos móveis de maneiras responsivas de maneira rápida e simples. (LIMA, 2022)

Originalmente, o *Bootstrap* havia sido desenvolvido para o *Twitter* por um grupo de desenvolvedores liderados por Mark Otto e Jacob Thornton Logo, tanto que antes de se tornar uma estrutura de código-aberto seu nome era “*Twitter Blueprint*”, após isso esta ferramenta se tornou uma das estruturas de *front-end* mais populares do mundo. (LIMA, 2022)

2.5.9 MVC

O modelo de arquitetura MVC (Model-View-Control) surgiu na década de 80, porém sua popularização se tornou mais evidente a partir do momento em quem aplicações WEB foram criadas.

Sua dinâmica funciona de forma simples e dinâmica possibilitando a divisão do projeto em camadas bem definidas, todas as requisições da aplicação são direcionadas inicialmente da camada *controller* que é a camada que intermedia as requisições enviadas pelo *view* com as respostas fornecidas pelo *model*, após a coleta de todas as informações disponíveis na camada *model* que representa o modelo de negócio gerenciando e controlando a forma como os dados se comportam, passando para o *controller* e chegando por fim ao *view* que é responsável por apresentar as informações de forma visual ao usuário.

3 Requisitos Do Sistema De Software

Este capítulo tem como objetivo analisar, detalhar e propor uma solução geral do sistema, sob o ponto de vista de negócio, de acordo com os requisitos levantados e validados no capítulo 3.

3.1 Requisitos Funcionais

Neste item devem ser descritos os requisitos funcionais que especificam ações que um sistema deve ser capaz de executar, ou seja, os objetivos do sistema, incluindo prioridade e regras de negócio. A seguir são apresentados exemplos.

[RF001] – Manter usuário

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: Este requisito permite que o usuário, possa fazer todas as operações de um CRUD.

[RF002] – Logar usuário

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: Este requisito permite que o usuário, possa fazer acessar sua conta cadastrada.

[RF003] – Inserir notícia

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: Este requisito permite que o usuário, escreva uma notícia em um aplicativo *Web* para verificar se é *fake* ou não.

[RF004] – Consultar notícias cadastradas

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: Com esse requisito o usuário consegue visualizar todas as notícias que ele já verificou.

[RF005] – Verificar notícia

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: Após o usuário digitar a notícia que quer verificar a veracidade, o sistema, apresentará se a notícia é *fake* ou não.

3.2 Requisitos Não-Funcionais

[RNF001] – WEB

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: O sistema pode ser acessado por qualquer dispositivo conectado à internet, permitindo assim a interação do usuário com o sistema.

[RNF002] – Linguagem de Programação

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: A linguagem usada será o *Python*, pois, é uma das linguagens mais apropriadas para inteligência artificial.

[RNF003] – Bibliotecas

Prioridade: ■ Essencial 📌 Importante 📌 Desejável

Descrição: O sistema será criado a partir das bibliotecas: *NLTK*, que será usada para o processamento dos textos. *NUMPY*, será usada para funções matemáticas. *PANDAS*, que fara análises gráficas. *SKLEARNING*, que será responsável pelo *Machine Learning* em si. *CSV*, para importar arquivos csv's.

LOGISTIC REGRESSION, que contém o algoritmo para detecção das *fake News*.
FLASK, para montar a aplicação web.

[RNF004] – Front end.

Prioridade: ■ Essencial ■ Importante ☞ Desejável

Descrição: A aplicação também será desenvolvida em html e css, em relação ao aspecto visual, para melhor interação com o usuário.

[RNF005] – Usabilidade

Prioridade: ☞ Essencial ☞ Importante ☞ Desejável

Descrição: O sistema deve prover uma interface simples e intuitiva, de fácil navegação para facilitar o uso do mesmo por parte dos usuários.

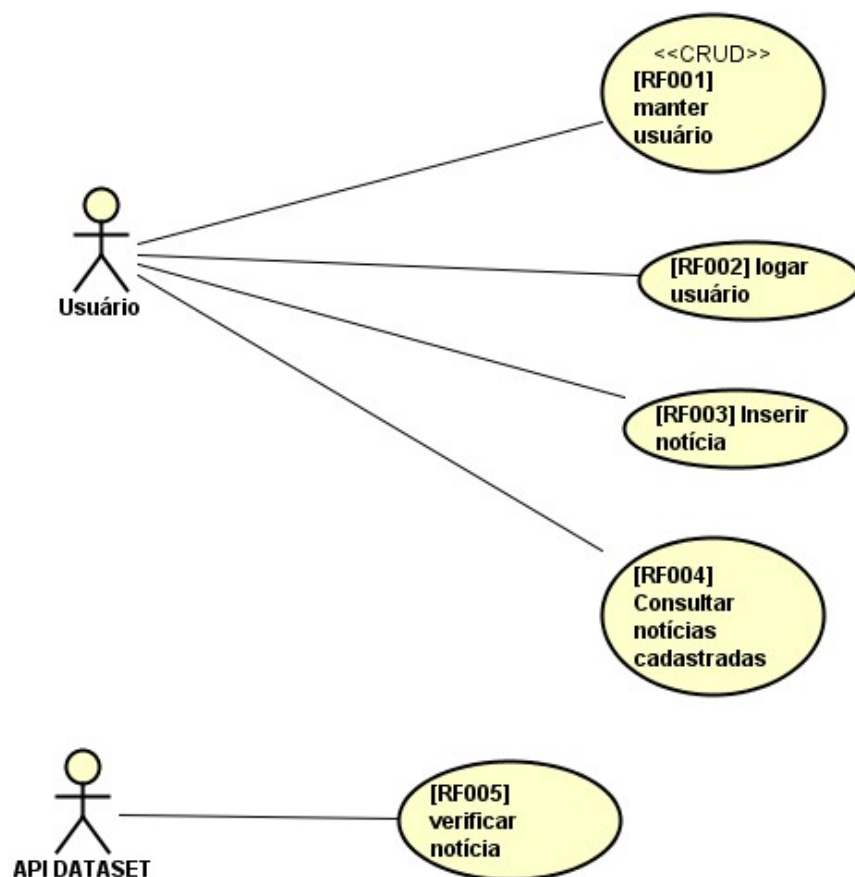
3.3 Modelagem Funcional

Neste item devem ser descritos os requisitos a serem atendidos funcionalmente pelo sistema de uma forma simples, possibilitando a compreensão do comportamento do sistema pela perspectiva do usuário. Devem ser descritos os atores e o diagrama de caso de uso. A seguir um exemplo de especificação de atores, do diagrama de caso de uso e da especificação de caso de uso.

3.3.1 Diagrama de Caso de Uso

A seguir é apresentada a notação básica de um diagrama de caso de uso

Figura 5 - Diagrama de Caso de Uso



Fonte: Autores 2022.

3.3.2 Atores

A seguir é apresentado um exemplo da especificação de atores.

Nome	Descrição
Usuário	Usuário do sistema responsável para verificação das notícias

3.3.3 Especificação do Caso de Uso

A seguir é apresentado um exemplo da especificação de casos de uso.

CSU01 - Efetuar Cadastro	
Sumário:	O usuário irá efetuar seu cadastro através tela de cadastro.
Ator principal:	Usuário
Ator secundário	Não se aplica
Pré-condições:	
Fluxo Principal: <ol style="list-style-type: none"> 1. O usuário irá digitar seu nome, e-mail e senha; 2. O sistema irá verificar se o e-mail e a senha, atende aos padrões solicitados; 3. O sistema irá verificar se o e-mail já existe; 4. Armazenar no banco de dados. 	
Fluxo alternativo – consultar dados usuário: <ol style="list-style-type: none"> 1. O usuário irá clicar no seu perfil; 2. O sistema apresentará seus dados e notícias que o usuário já verificou com seus resultados. 	
Fluxo alternativo – atualizar dados usuário: <ol style="list-style-type: none"> 1. O usuário irá clicar em atualizar perfil; 2. O sistema apresentará o campo de atualização de nome e senha; 	
Fluxo alternativo – atualizar dados usuário <ol style="list-style-type: none"> 1. O usuário irá clicar em deletar perfil; 2. O irá solicitar que digite sua senha; 3. O sistema mostrará uma janela perguntando se realmente deseja deletar o perfil; 4. Clicando em ok, o cadastro será removido do banco de dados 	
Fluxo de exceção cadastro – campo faltante a ser digitado, e-mail já existente e senha fora dos padrões <ol style="list-style-type: none"> 1. O sistema irá apresentar uma mensagem no campo faltante, pedindo a ser digitado 2. Se o e-mail não atender as especificações padrão, o usuário deve digitar novamente um e-mail válido; 3. Se a senha não atender aos padrões (8 caracteres, pelo menos 1 letra 	

<p>maiúscula e pelo menos 1 número), irá apresentar uma mensagem solicitando digitar novamente a senha;</p> <p>4. Caso seja digitado um e-mail já existente no cadastro, apresentará uma mensagem solicitando um novo e-mail não cadastrado.</p>
<p>Fluxo de exceção atualizar – senha não segue os padrões:</p> <ol style="list-style-type: none"> 1. O sistema apresentará uma mensagem que a senha segue os padrões solicitados (8 caracteres, pelo menos 1 letra maiúscula e pelo menos 1 número); 2. E solicitará que digite a senha novamente
<p>Fluxo de exceção deletar – senha não corresponde ao perfil:</p> <ol style="list-style-type: none"> 1. O sistema apresentará uma mensagem que a senha está incorreta e não poderá deletar o perfil; 2. E solicitará que digite novamente

CSU02 – Efetuar Login	
Sumário:	O usuário digitará seus dados cadastro para efetuar seu login.
Ator principal:	Usuário
Ator secundário	Não se aplica
Pré-condições: Acesso local	
<p>Fluxo Principal:</p> <ol style="list-style-type: none"> 1. O usuário irá digitar seu e-mail e senha cadastrados; 2. O sistema irá verificar se eles estão cadastrados; 3. Após a verificação, sistema irá habilitar a página de verificação. 	
<p>Fluxo de exceção – Usuário não cadastrado</p> <ol style="list-style-type: none"> 1. O sistema irá apresentar uma mensagem que não existe senha ou e-mail, cadastrado no sistema. 	
CSU03 - ESCREVER NOTÍCIA	
Sumário:	O usuário escreverá a notícia na página web para que

	possa verificar sua autenticidade.
Ator principal:	Usuário
Ator secundário	Não se aplica
Pré-condições: Acesso local	
Fluxo Principal: <ol style="list-style-type: none"> 1. O usuário irá digitar o título e notícia inteira no campo adequado; 2. O sistema irá verificar se os campos de título e notícia foram digitados; 3. Após a verificação, sistema irá habilitar a verificação; 4. Armazenar no banco de dados. 	
Fluxo de exceção – campo faltante a ser digitado. <ol style="list-style-type: none"> 1. O sistema irá apresentar uma mensagem no campo faltante, pedindo ser digitado. 	

CSU04 – Verificar notícia	
Sumario:	Notícia digitada será verificada sua veracidade
Ator principal:	Sistema
Ator secundário:	Não se aplica
Pré-condições: Acesso local	
Fluxo Principal: <ol style="list-style-type: none"> 1. O sistema irá verificar se a notícia tem pelo menos 100 palavras e se encontra em língua portuguesa; 2. A notícia será analisada por meio do aprendizado automático (capítulo 5.2), reconhecimento seus padrões; 3. O sistema apresentará mensagem informando se a notícia é verdadeira ou falsa. 	
Fluxo de exceção - notícia não atende requisitos mínimos <ol style="list-style-type: none"> 1. O sistema apresentará uma mensagem uma mensagem que ele se encontra em outra língua ou que não possui o tamanho mínimo de caracteres. 2. Retornar ao passo 1 do principal. 	

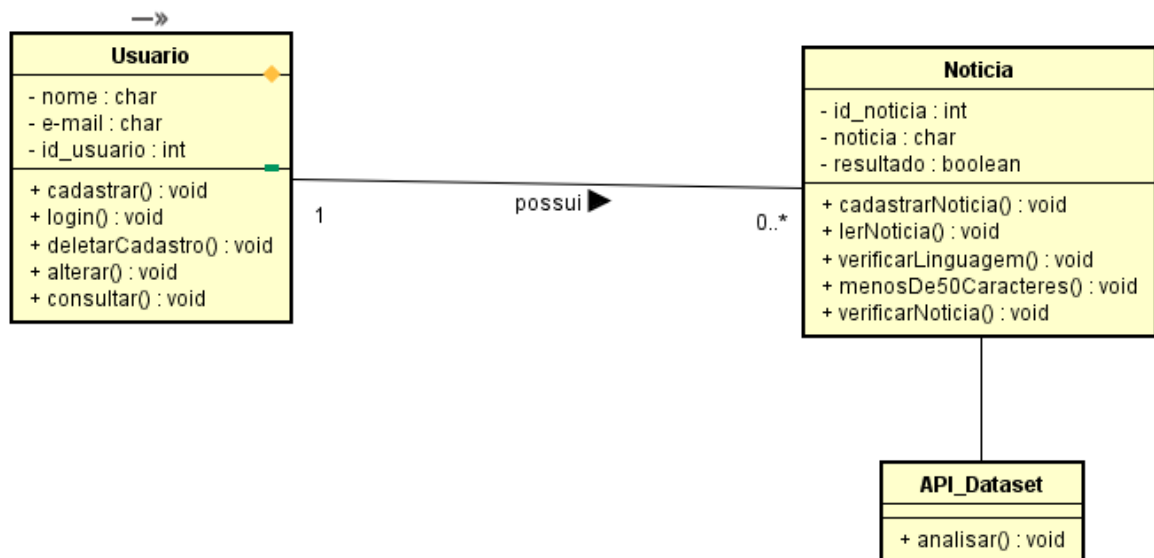
4 Análise

Este capítulo tem como objetivo analisar, detalhar e propor uma solução geral do sistema, sob o ponto de vista de negócio, de acordo com os requisitos levantados e validados no capítulo 3. Além disso, é apresentado o refinamento da proposta de solução geral do sistema, apresentando a solução técnica, incluindo a visão de projeto e implementação, a arquitetura e a tecnologia utilizada.

4.1 Diagrama de Classes de Análise (Visão de Negócio)

A seguir é apresentada a notação básica de um diagrama de classes.

Figura 6 - Diagrama de Classes

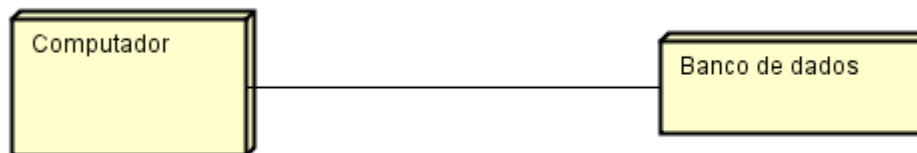


Fonte: Autores 2022.

5 Projeto

5.1 Arquitetura do Sistema

Figura 7 - Diagrama de Implementação



Fonte: Autores 2022

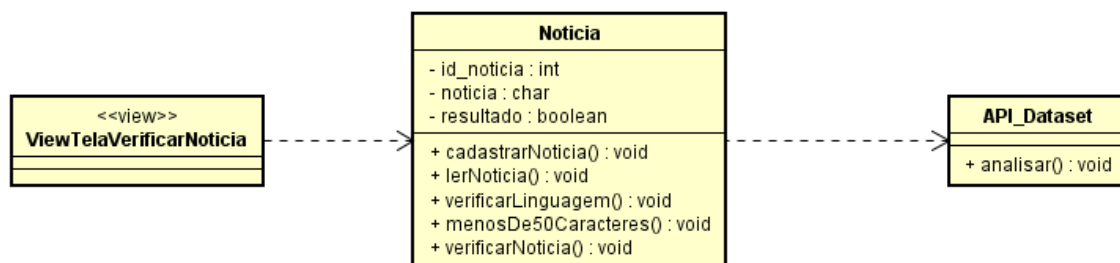
5.2 Diagrama de Classes de Projeto por Caso de Uso

Para a análise das notícias primeiro precisa transformar as palavras em números, para que os algoritmos testados possam funcionar. Primeiro será feito o pré-processamento do texto, retirando *stopwords*, pontuações e stemitização. Por último no pré-processamento, é usado a função *TfidfTransformer* e *Bag of Words*, que terá o objetivo de calcular o peso de importância que cada palavra tem na notícia. Depois da notícia ter passado por esse pré-processamento, ela será jogada para os algoritmos de predição que fará análise e dirá se a notícia é verdadeira ou não.

Foram usados os algoritmos de SVM, Regressão Logística, MLPClassifier e Naive Bayes (Bernoulli e Multinomial) com base em testes, foram escolhidos os 3 melhores resultados, que foram o SVM, Regressão logística, MLPClassifier, para que assim fique uma predição mais confiável.

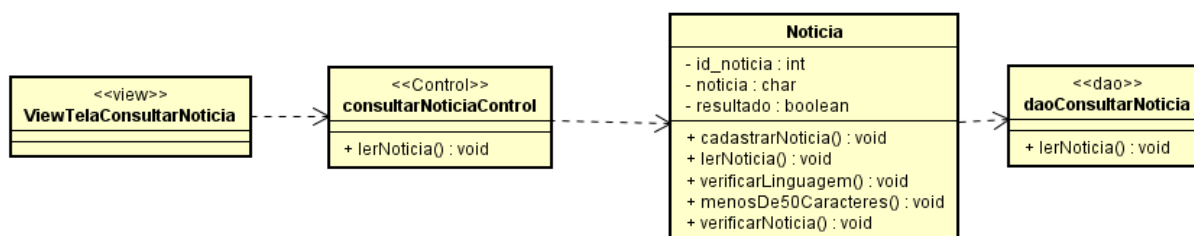
A partir desses modelos é possível calcular ou prever probabilidade de um evento ocorrer.

Figura 8 - Diagrama de Caso de Uso - Detecção da veracidade



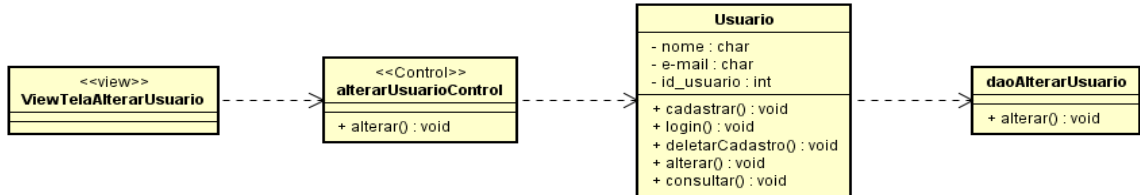
Fonte: Autores 2022

Figura 9 - Diagrama de Caso de Uso - Consultar notícias já verificadas



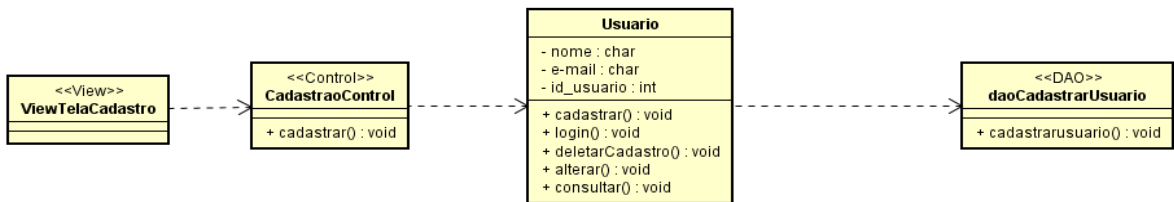
Fonte: Autores 2022

Figura 10 - Diagrama de Caso de Uso - Alterar Usuário



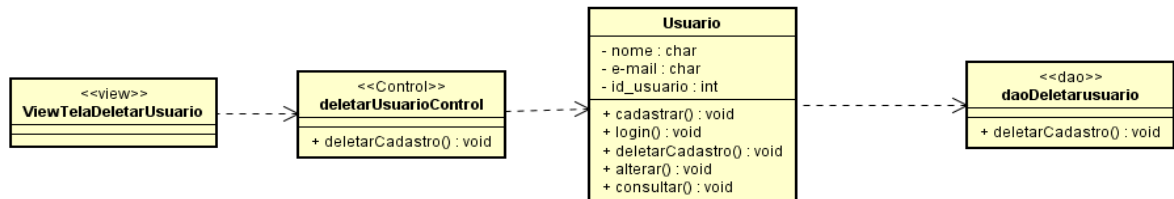
Fonte: Autores 2022

Figura 11 - Diagrama de Caso de Uso - Cadastrar usuário



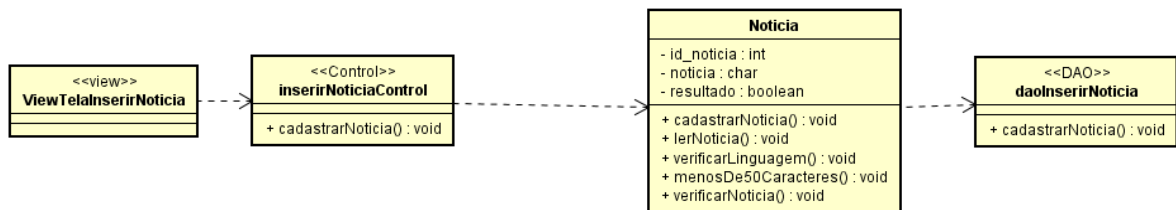
Fonte: Autores 2022

Figura 12 - Diagrama de Caso de Uso - Deletar usuário



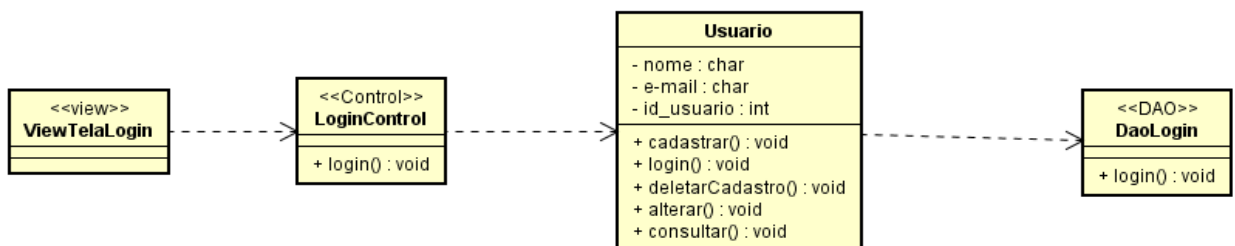
Fonte: Autores 2022

Figura 13 - Diagrama de Caso de Uso - Inserir notícia no bando de dados



Fonte: Autores 2022

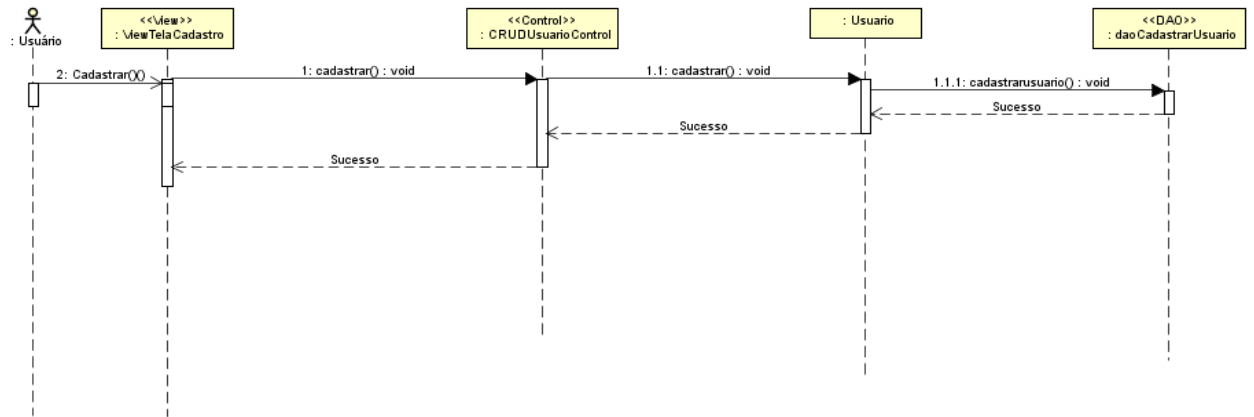
Figura 14 - Diagrama de Caso de Uso - Login



Fonte: Autores 2022

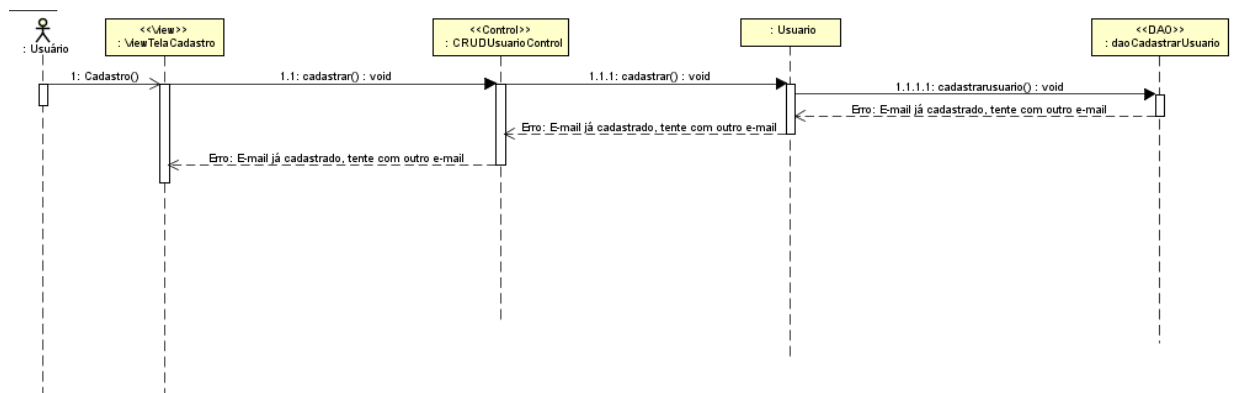
5.3 Diagrama de Sequências

Figura 15 - Diagrama de Sequências - Cadastro com sucesso



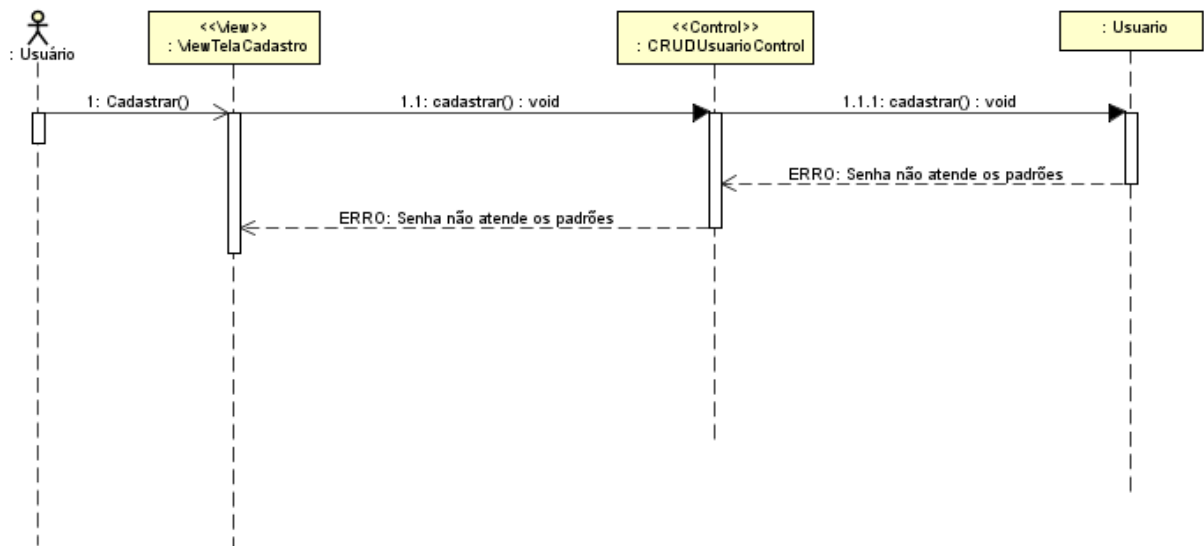
Fonte: Autores 2022

Figura 16 - Diagrama de Caso de Uso - Cadastro e-mail já cadastrado



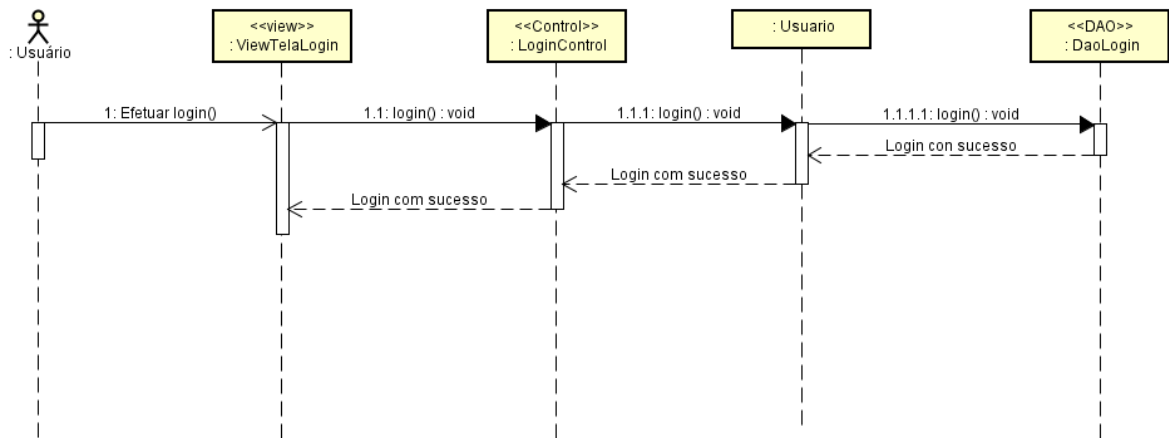
Fonte: Autores 2022

Figura 17 - Diagrama de Caso de Uso - Cadastrar erro na senha



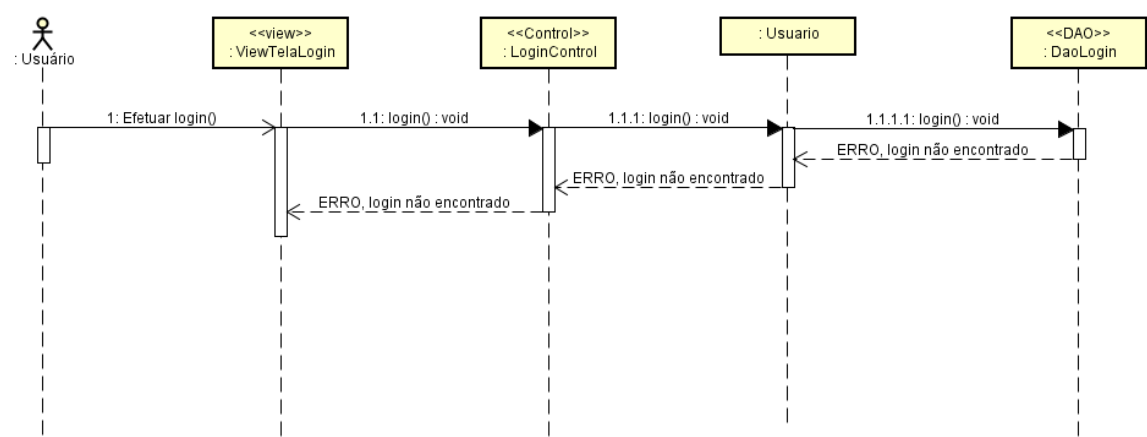
Fonte: Autores 2022

Figura 18 - Diagrama de Caso de Uso - Login



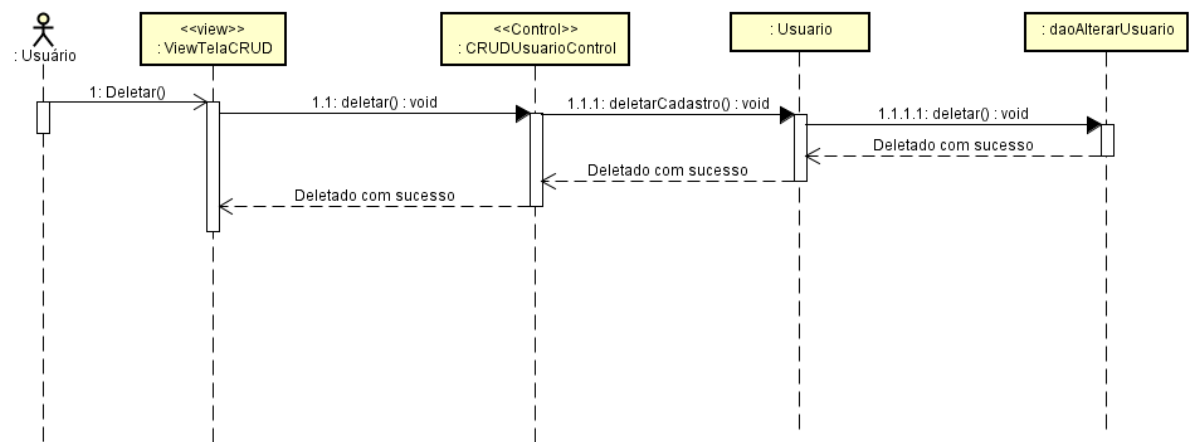
Fonte: Autores 2022

Figura 19 - Diagrama de Caso de Uso - Login não encontrado



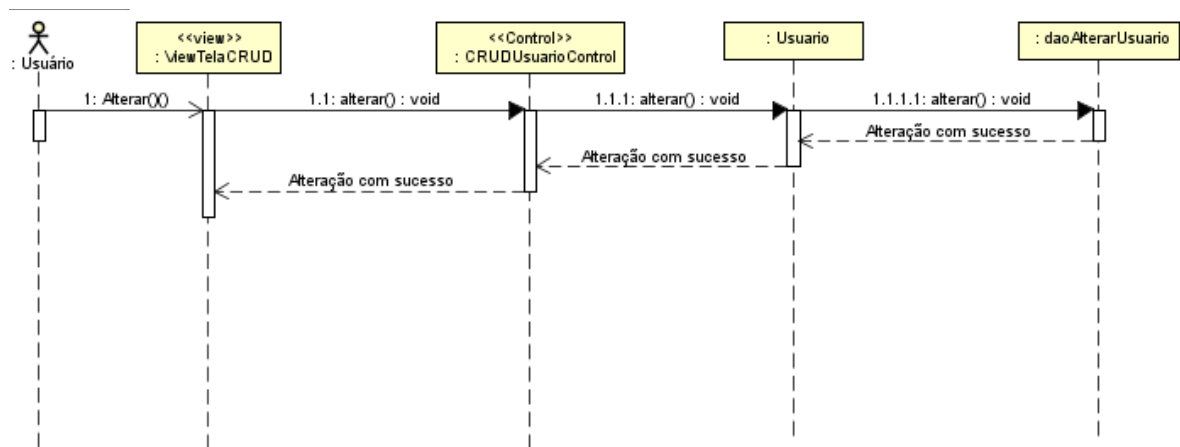
Fonte: Autores 2022

Figura 20 - Diagrama de Caso de Uso - Deletar



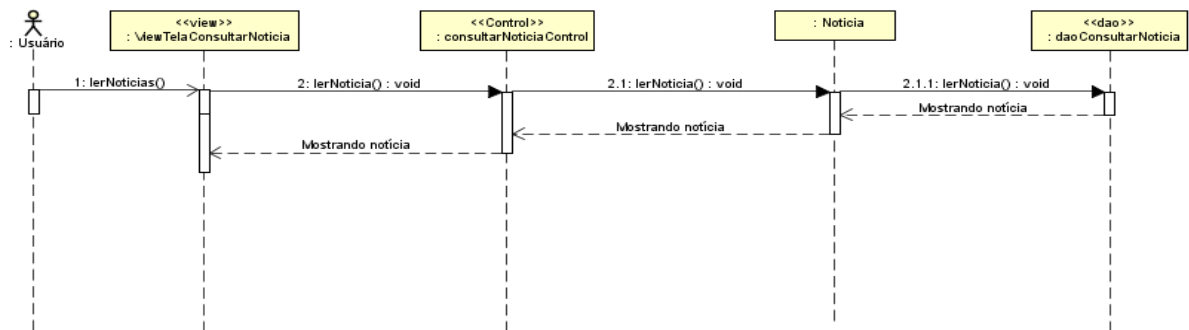
Fonte: Autores 2022

Figura 21 - Diagrama de Caso de Uso - Alterar



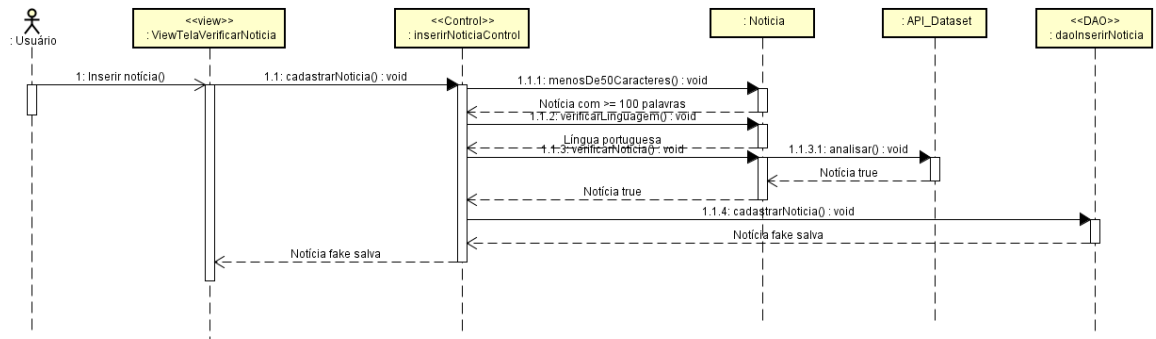
Fonte: Autores 2022

Figura 22 - Diagrama de Caso de Uso - Visualizar notícias cadastradas



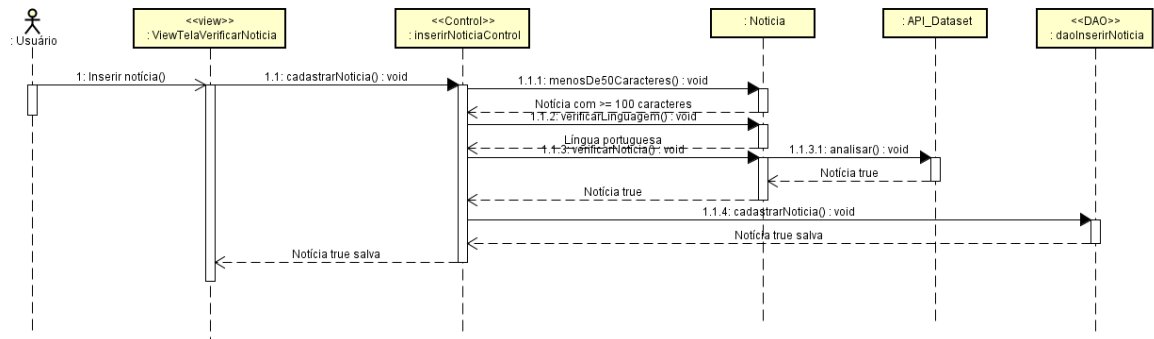
Fonte: Autores 2022

Figura 23 - Diagrama de Caso de Uso - Notícia falsa



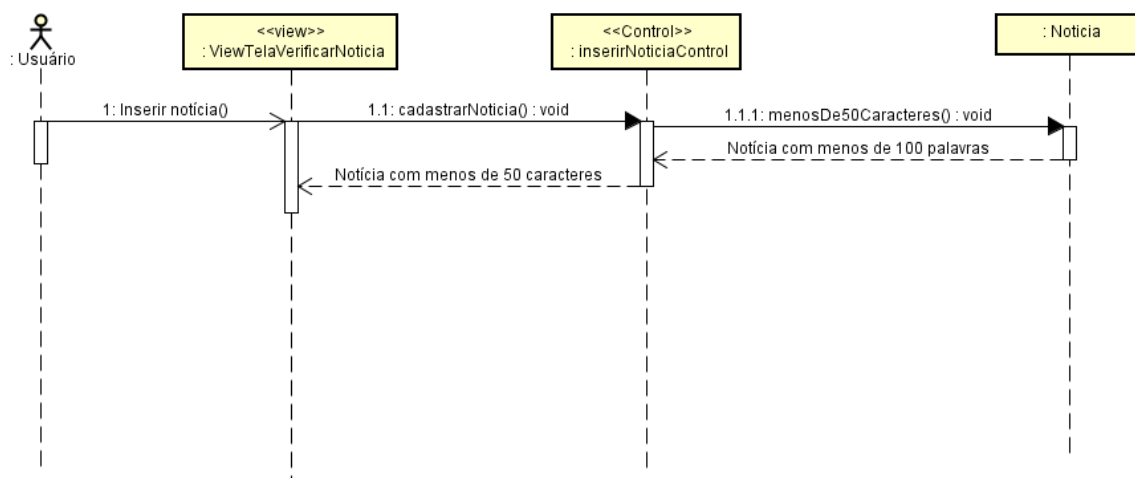
Fonte: Autores 2022

Figura 24 - Diagrama de Caso de Uso - Notícia verdadeira



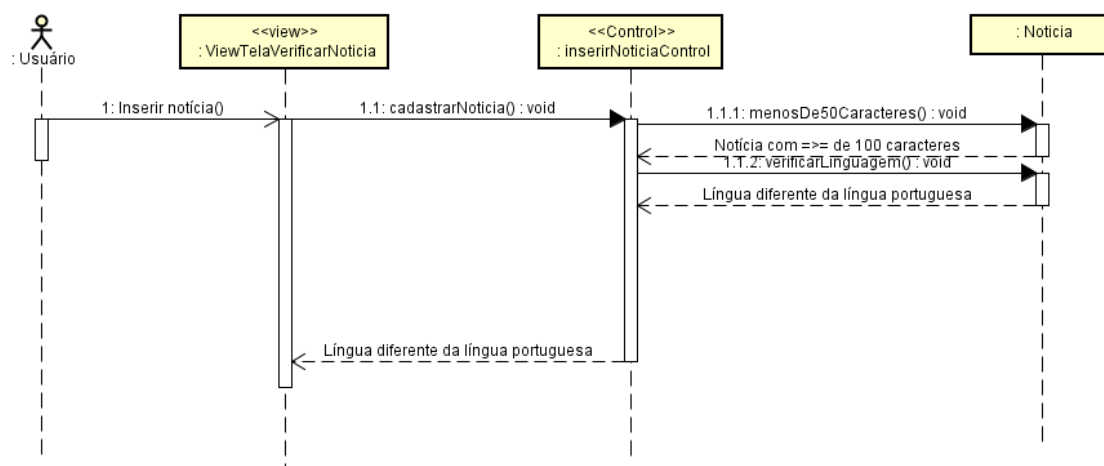
Fonte: Autores 2022

Figura 25 - Diagrama de Caso de Uso - Quantia de palavras não OK



Fonte: Autores 2022

Figura 26 - Diagrama de Caso de Uso - Língua Portuguesa não OK

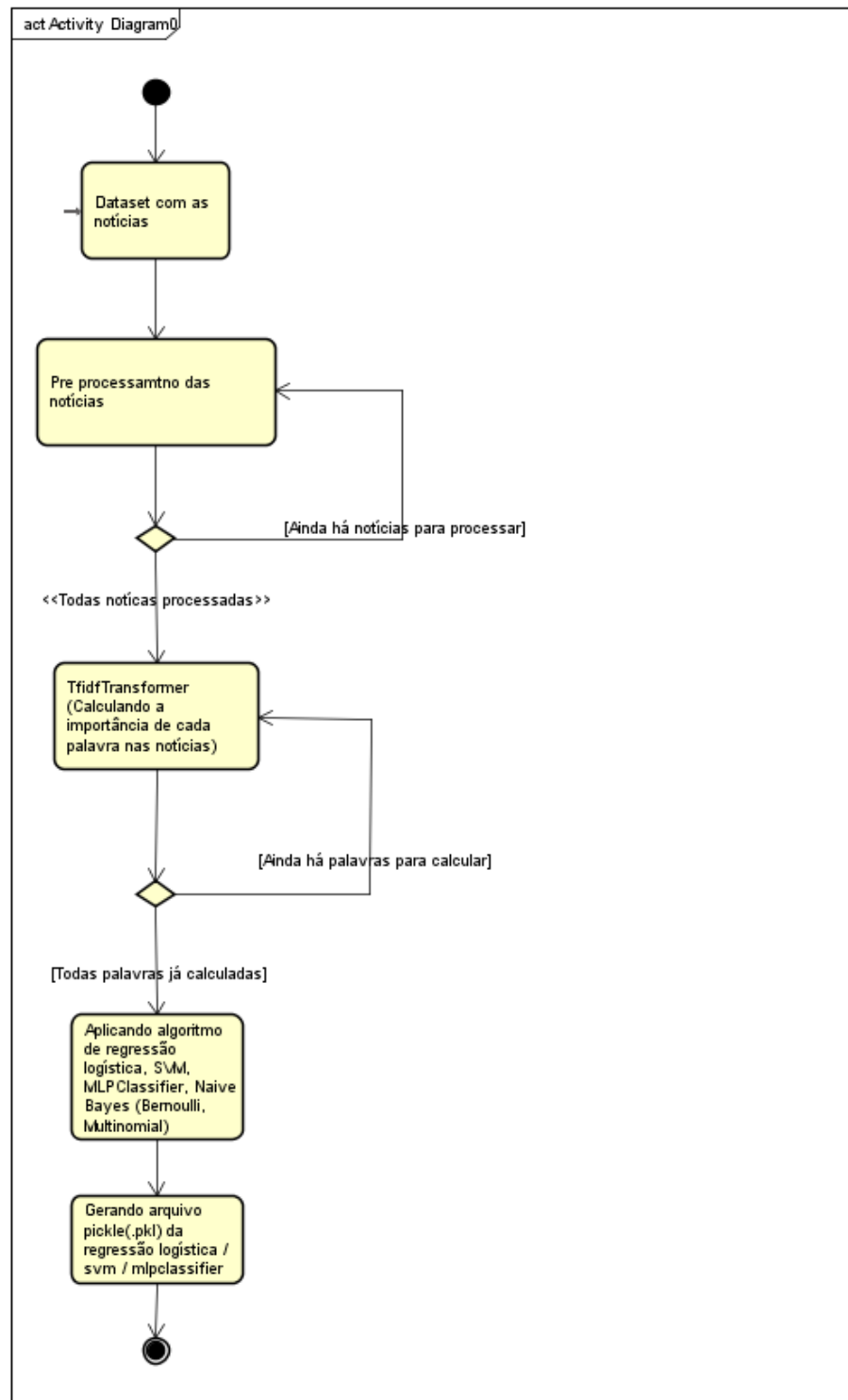


Fonte: Autores 2022

5.4 Diagrama de Atividades

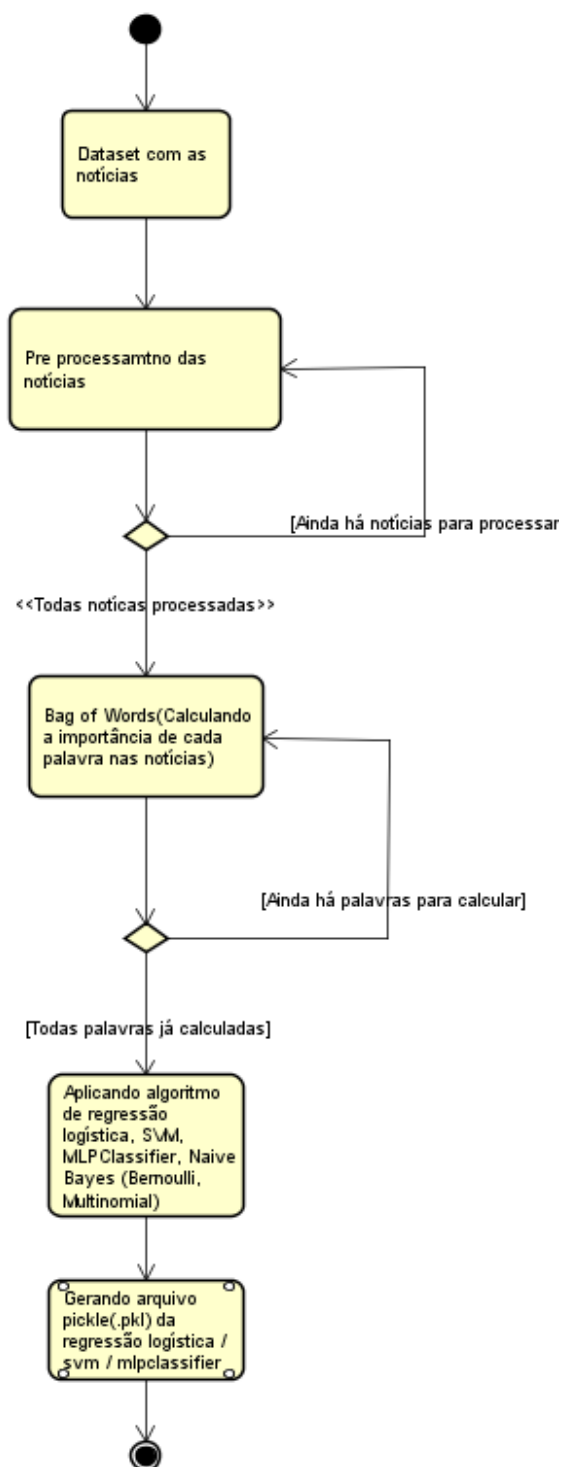
O diagrama de atividades representa o detalhamento de tarefas e o fluxo de uma atividade para outra de um sistema, geralmente utilizado para os métodos que contém regras de negócio. A seguir é apresentada a notação básica de um diagrama de atividades. Esse diagrama somente deverá ser elaborado se houver necessidade e agregar valor ao projeto.

Figura 27 - Diagrama de Atividades – Geração do Algoritmo com TF-IDF



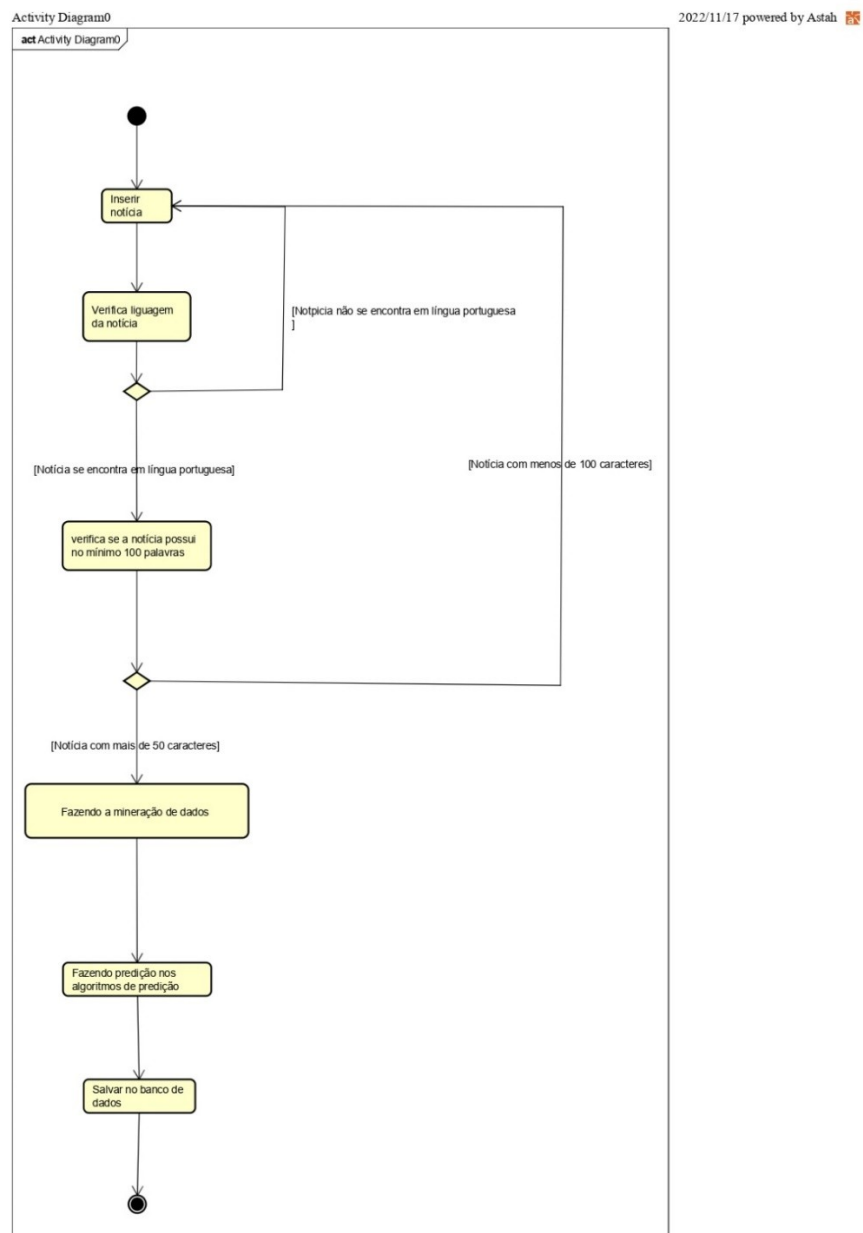
Fonte: Autores 2022

Figura 28 - Diagrama de Atividades – Geração do Algoritmo com Bag of Words



Fonte: Autores 2022

Figura 29 - Diagrama de Atividades – Atividade da detecção e gravação no bando de dados

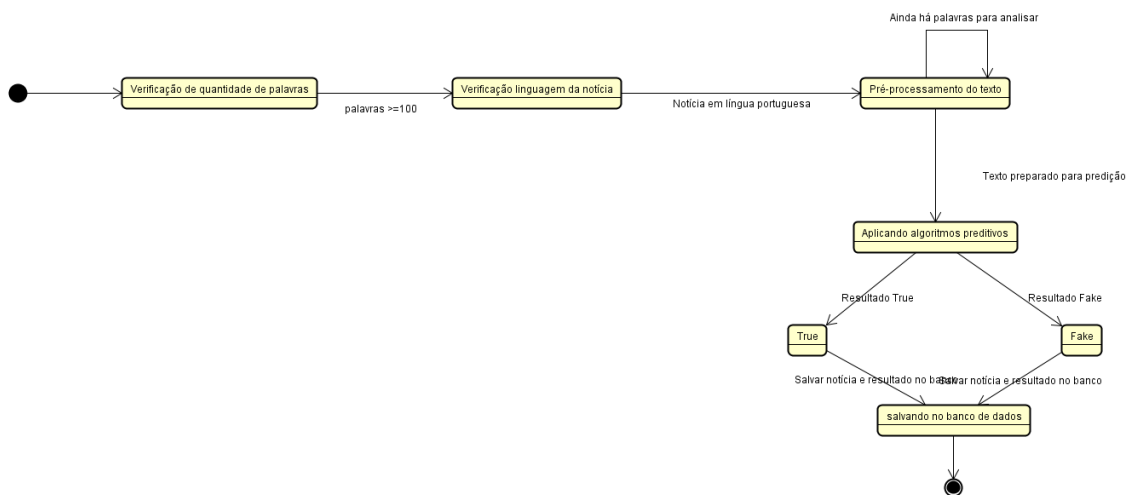


Fonte: Autores 2022

5.5 Diagrama de Estados

O diagrama de estados especifica as sequências de estados pelas quais o objeto pode passar durante seu ciclo de vida em resposta a eventos. A seguir é apresentada a notação básica de um diagrama de estados. **Esse diagrama somente deverá ser elaborado se houver necessidade e agregar valor ao projeto.**

Figura 30 - Diagrama de Estados – Atividade da detecção e gravação no bando de dados



Fonte: Autores 2022

6 Testes

Este capítulo tem como objetivo identificar erros no sistema, validar as funções do sistema, verificar se os requisitos foram implementados de forma adequada. Sugere-se a criação de um plano de testes e um roteiro de testes baseado nos casos de uso. Também pode-se utilizar a técnica de TDD (*Test Driven Development*), neste caso os testes também devem ser registrados.

6.1 Plano de Testes

Finalidade

Este Plano de Teste referente ao Sistema Apresentação atende aos seguintes objetivos:

1. Identifica os itens que devem ser inspecionados pelos testes.
2. Identifica a motivação e as ideias subjacentes às áreas de teste a serem abrangidas.
3. Descreve a abordagem de teste que será usada.
4. Identifica os recursos necessários e fornece uma estimativa dos esforços de teste.
5. Lista os elementos liberados do projeto de teste.

6.1.1 Escopo

Este plano de teste abordará testes de unidade e de sistema do Verificador de Fake News com aprendizado de máquina– Módulo 1) Neste Módulo ainda não há integração com o Sistema de Expedição e, portanto, esta integração não será testada.

6.2 Itens-alvo dos testes

A tabela abaixo lista os casos de testes, baseado nos casos de uso, que serão sujeitos a testes funcionais e a priorização de cada um deles:

Item-alvo	Fator de Risco (Impacto)	Ordem de Prioridade
Inserir notícia	3	1º.
Verificar se é falsa ou verdadeira	5	2º.
Voltar e verificar outra notícia	2	3º.
Cadastro CRUD	1	4º

6.3 Resumo dos testes planejados

6.3.1 Resumo das inclusões dos testes

Os principais testes planejados para o Módulo 1 são:

- Teste funcional do caso de uso “Inserir Notícia”, para verificar se é possível inserir a notícia.
- Teste funcional de todo o caso de uso “Verificar Notícia”;
- Teste funcional do caso de uso “Voltar e verificar outra notícia”.
- Teste funcional no caso de uso “CRUD”

6.3.2 Resumo dos outros candidatos a possível inclusão

A seguir temos um resumo de áreas de teste cuja avaliação e investigação poderão ser úteis, mas que ainda não foram suficientemente pesquisadas, e que a princípio não serão priorizados para testes:

- Testes de Desempenho;

6.3.3 Necessidades Ambientais

Os conjuntos de tabelas a seguir apresentam os recursos do sistema necessários ao esforço de teste descrito neste Plano de Teste.

Recursos do Sistema		
Recurso	Quantidade	Nome e Tipo
Rede Local	1	Rede Local do Usuário
PCs de Teste	1	Estação de trabalho com navegador e acesso à Intranet

6.3.4 Elementos de softwares básicos do ambiente de teste

Nome do Elemento de Software	Tipo e Outras Observações
Windows 10	Sistema Operacional
Chrome 94.X	Navegador da Internet

6.4 Responsabilidades, perfil da equipe e necessidades de treinamento

6.4.1 Hardware básico do sistema

Recursos Humanos		
Papel	Recursos Mínimos Recomendáveis (número de papéis alocados em tempo integral)	Responsabilidades ou Comentários Específicos
Analista de Teste	1	Identifica e define os testes específicos a serem conduzidos e quais ferramentas de testes serão utilizadas e como.
Testador	1	Implementa e executa os testes.

O risco mais evidente na execução deste Plano de Teste é a falta de conhecimento do usuário, ao digitar uma notícia.

A não execução dos testes de desempenho pode levar a uma má experiência do usuário com o Hardware básico do sistema.

7 RESULTADOS

Nesse capítulo irá ser apresentado os resultados encontrados e como a aplicação está funcionando.

7.1 Dataset

No projeto é usado um dataset com 7.200 notícias em língua portuguesa-BR, das quais 3600 notícias verdadeiras e 3.600 notícias. Ele possui 3 colunas, uma de index (numera as notícias), *label* (identifica se a notícia é falsa ou não) e 'preprocessed_news' (nela está a notícia inteira):

Figura 31 - Dataset

index	label	preprocessed_news
0	fake	katia abreu diz vai colocar expulsao moldura n...
1	fake	ray peita bolsonaro conservador fake entrevist...
2	fake	reinaldo azevedo desmascarado policia federal ...
3	fake	relatorio assustador bndes mostra dinheiro pub...
4	fake	radialista americano fala sobre pt vendem ilus...
...
7195	true	jornal britanico acao contra lula lava jato se...
7196	true	temer diz acionou pf cade investigar aumentos ...
7197	true	obstaculos politicos temer especialistas ouvid...
7198	true	setembro boa noite aqui estao principais notic...
7199	true	envolve politica diz brasileiro preso venezuel...

Fonte: Autores 2022

Esse dataset pode ser encontrado no link:
<https://github.com/roneysco/Fake.br-Corpus>

7.2 Pré-processamento

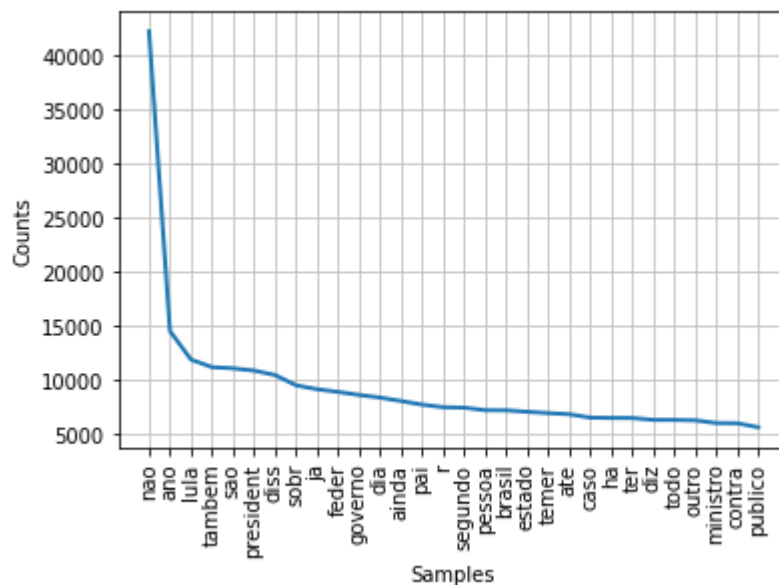
O pré-processamento consisti em retirar dados que não sejam relevantes ao aprendizado de máquina e faça com que seu aprendizado seja mais rápido.

Primeiro foi retirado qualquer caractere que fosse diferente de palavras, assim foram retirados pontos, números e caracteres especiais, logo em seguida foram retiradas as *stopwords* dos textos, e por último foi feita a stemização do texto.

7.3 Palavras mais frequentes no dataset por inteiro

Para essa análise foi contabilizado a quantidade de vezes que cada palavra aparece nas 7.200 notícias. E com o resultado foi feito uma nuvem de palavras e um gráfico das 30 palavras com maior frequência.

Figura 32 - Gráfico com as 30 palavras com maior frequência:



Fonte: Autores 2022

Figura 37 - Nuvem de palavras das notícias verdadeiras

Fonte: Autores 2022

7.4 Resultados com a técnica TF-IDF

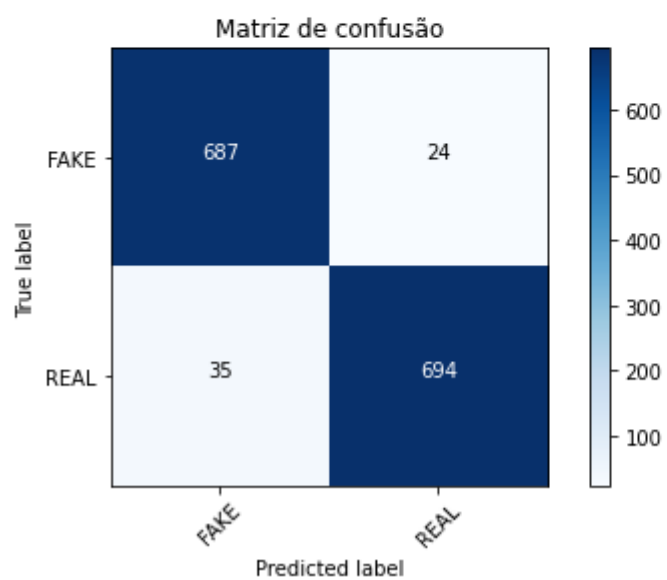
Utilizando a técnica *TF-IDF*, foram testados alguns algoritmos de predição que foram regressão logística, *SVM*, *Naive Bayes* (*Bernoulli* e *Multinomial*) e todos tiveram um bom resultado. Segue os resultados obtidos.

Figura 38 - Resultado da regressão logística com a matriz de confusão

Acurácia da classificação da regressão logística: 95.9%

Classification Report da classificação da regressão logística:

	precision	recall	f1-score	support
fake	0.95	0.97	0.96	711
true	0.97	0.95	0.96	729
accuracy			0.96	1440
macro avg	0.96	0.96	0.96	1440
weighted avg	0.96	0.96	0.96	1440



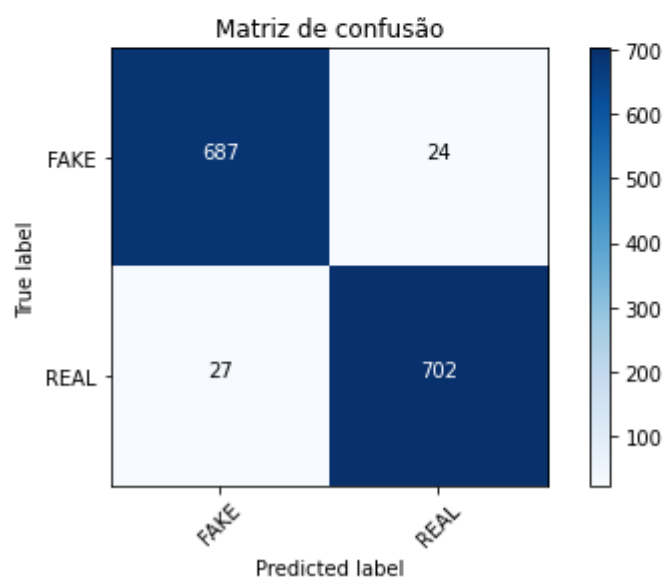
Fonte: Autores 2022

Figura 39 - Resultado da SVM com a matriz de confusão

Acurácia da classificação do SVM: 96.46%

Classification Report da classificação do SVM:

	precision	recall	f1-score	support
fake	0.96	0.97	0.96	711
true	0.97	0.96	0.96	729
accuracy			0.96	1440
macro avg	0.96	0.96	0.96	1440
weighted avg	0.96	0.96	0.96	1440



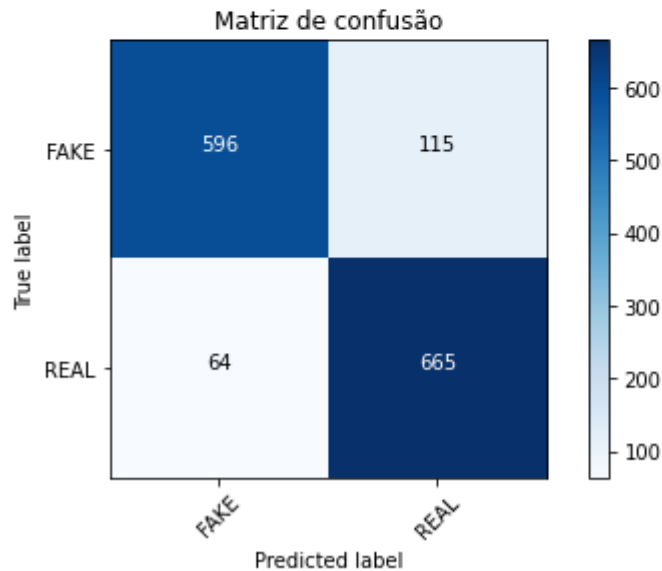
Fonte: Autores 2022

Figura 40 - Resultado Naive Bayes MultinomialNB com a matriz de confusão

Acurácia da classificação do MultinomialNB: 87.57%

Classification Report da classificação do MultinomialNB:

	precision	recall	f1-score	support
fake	0.90	0.84	0.87	711
true	0.85	0.91	0.88	729
accuracy			0.88	1440
macro avg	0.88	0.88	0.88	1440
weighted avg	0.88	0.88	0.88	1440



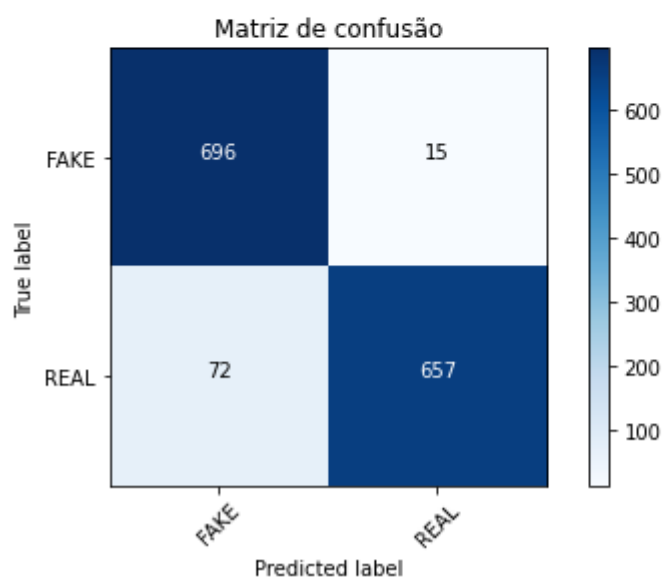
Fonte: Autores 2022

Figura 41 - Resultado Naive Bayes BernoulliNB com a matriz de confusão:

Acurácia da classificação do BernoulliNB: 93.96%

Classification Report da classificação do BernoulliNB:

	precision	recall	f1-score	support
fake	0.91	0.98	0.94	711
true	0.98	0.90	0.94	729
accuracy			0.94	1440
macro avg	0.94	0.94	0.94	1440
weighted avg	0.94	0.94	0.94	1440



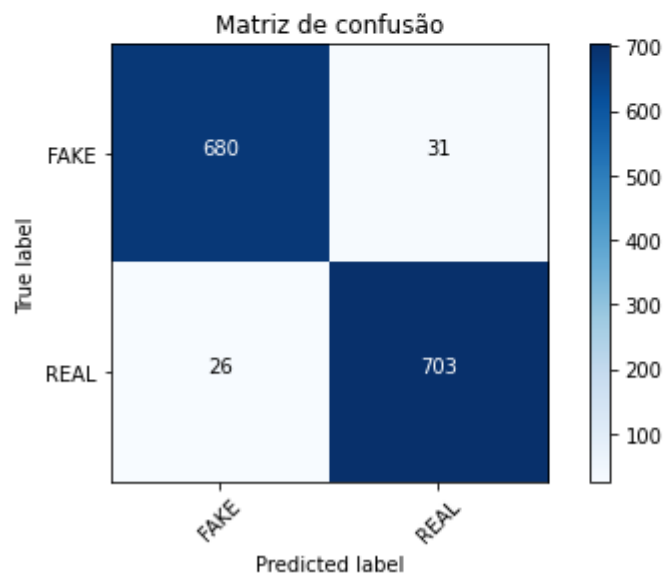
Fonte: Autores 2022

Figura 42 - Resultado MLPClassifier com a matriz de confusão:

Acurácia da classificação do MLPClassifier: 96.04%

Classification Report da classificação do MLPClassifier:

	precision	recall	f1-score	support
fake	0.96	0.96	0.96	711
true	0.96	0.96	0.96	729
accuracy			0.96	1440
macro avg	0.96	0.96	0.96	1440
weighted avg	0.96	0.96	0.96	1440



Fonte: Autores 2022

Após todos os testes, se percebe que todos algoritmos de predição tiveram ótimos resultados na acurácia, o *precision*, *recal* e o *f1-score* mantiveram coerência. O melhor resultado foi com o algoritmo *SVM* com resultado de 96,46% de acurácia e o mais baixo foi o de *Naive Bayes MultinomialNB* com 87,57 % de acurácia.

7.5 Resultados com a técnica Bag of Words

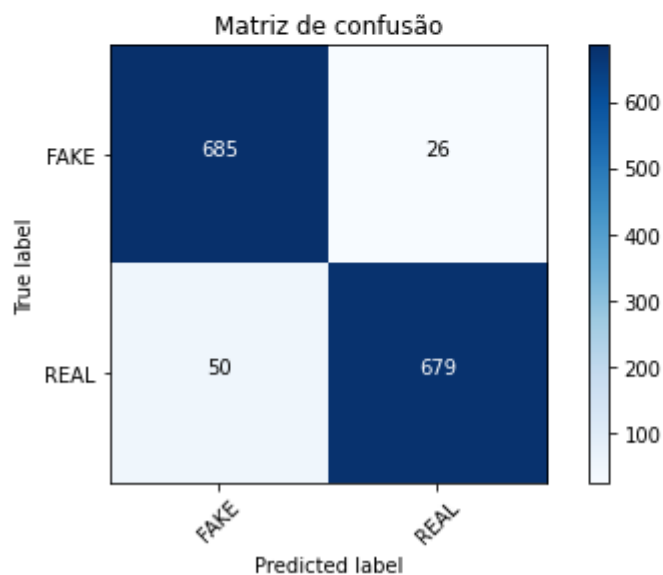
Utilizando a técnica *Bag of Words*, foram testados alguns algoritmos de predição que foram regressão logística, SVM, Naive Bayes (*Bernoulli* e *Multinomial*) e todos tiveram um bom resultado. Segue os resultados obtidos.

Figura 43 - Resultado da regressão logística com a matriz de confusão:

Acurácia da classificação da regressão logística: 94.72%

Classification Report da classificação da regressão logística:

	precision	recall	f1-score	support
fake	0.93	0.96	0.95	711
true	0.96	0.93	0.95	729
accuracy			0.95	1440
macro avg	0.95	0.95	0.95	1440
weighted avg	0.95	0.95	0.95	1440



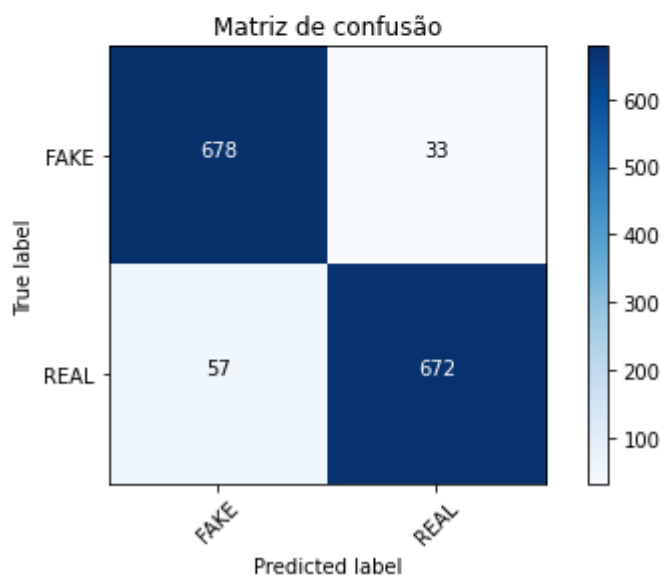
Fonte: Autores 2022

Figura 44 - Resultado da SVM com a matriz de confusão

Acurácia da classificação do SVM: 93.75%

Classification Report da classificação do SVM:

	precision	recall	f1-score	support
fake	0.92	0.95	0.94	711
true	0.95	0.92	0.94	729
accuracy			0.94	1440
macro avg	0.94	0.94	0.94	1440
weighted avg	0.94	0.94	0.94	1440



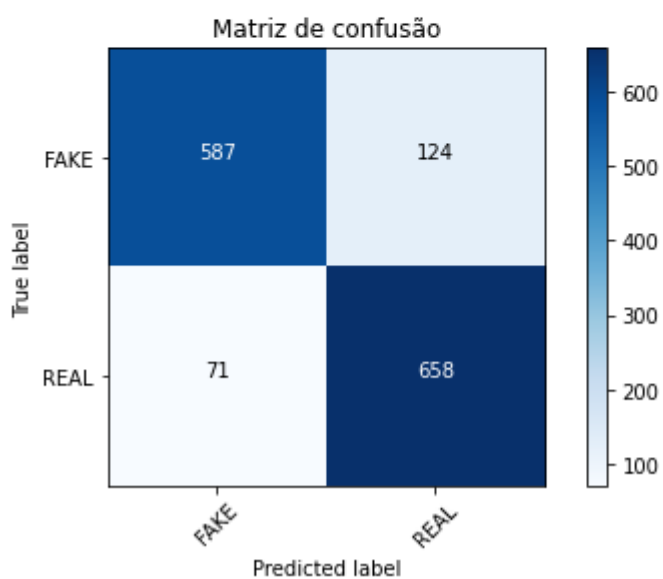
Fonte: Autores 2022

Figura 45 - Resultado Naive Bayes MultinomialNB com a matriz de confusão:

Acurácia da classificação do MultinomialNB: 86.46%

Classification Report da classificação do MultinomialNB:

	precision	recall	f1-score	support
fake	0.89	0.83	0.86	711
true	0.84	0.90	0.87	729
accuracy			0.86	1440
macro avg	0.87	0.86	0.86	1440
weighted avg	0.87	0.86	0.86	1440



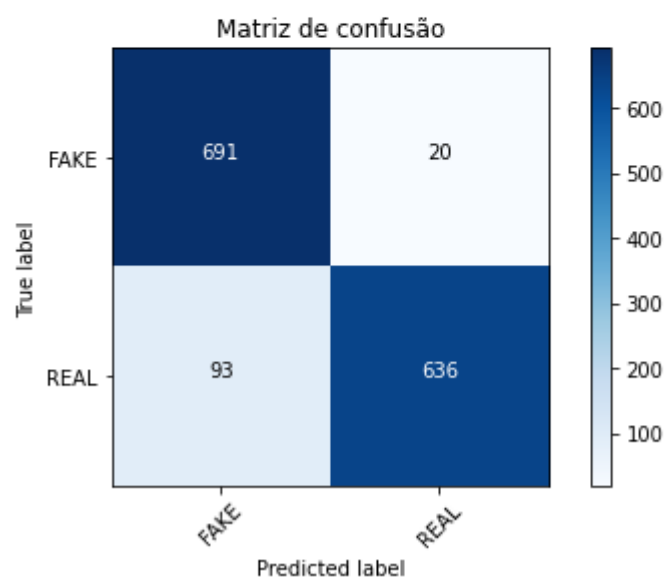
Fonte: Autores 2022

Figura 46 - Resultado Naive Bayes BernoulliNB com a matriz de confusão

Acurácia da classificação do BernoulliNB: 92.15%

Classification Report da classificação do BernoulliNB:

	precision	recall	f1-score	support
fake	0.88	0.97	0.92	711
true	0.97	0.87	0.92	729
accuracy			0.92	1440
macro avg	0.93	0.92	0.92	1440
weighted avg	0.93	0.92	0.92	1440



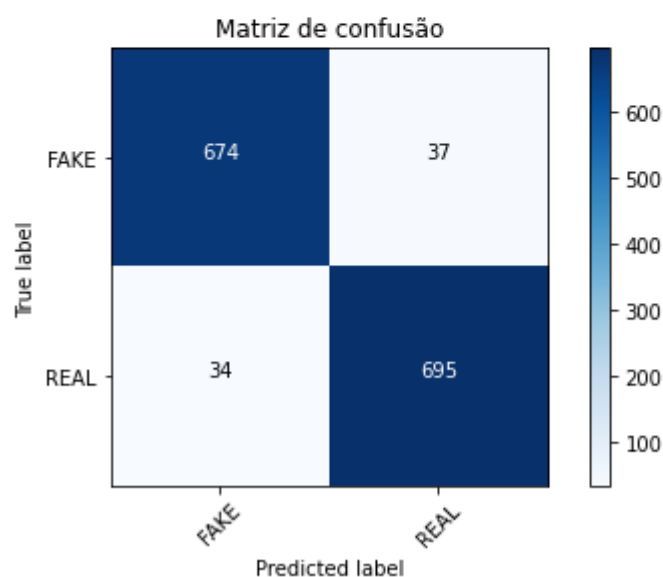
Fonte: Autores 2022

Figura 47 - Resultado MLPClassifier com a matriz de confusão:

Acurácia da classificação do MLPClassifier: 95.07%

Classification Report da classificação do MLPClassifier:

	precision	recall	f1-score	support
fake	0.95	0.95	0.95	711
true	0.95	0.95	0.95	729
accuracy			0.95	1440
macro avg	0.95	0.95	0.95	1440
weighted avg	0.95	0.95	0.95	1440



Fonte: Autores 2022

Após todos os testes, se percebe que todos algoritmos de predição tiveram ótimos resultados na acurácia, e o *precision*, *recall* e *f1-score* mantiveram coerência. O melhor resultado foi com o algoritmo de regressão logística com resultado de 94,72% de acurácia e o mais baixo foi o de Naive Bayes MultinomialNB com 86,46 % de acurácia.

7.6 Salvando o aprendizado dos 3 melhores resultados em arquivo.pkl

Após os testes se percebe que os melhores resultados foram usando *TF-IDF* com o algoritmo *SVM*, *Regressão logística* e *MLPClassifier*. Com esses resultados foi salvo o aprendizado desse algoritmo em um 'arquivo.pkl' que será usado na aplicação do site.

7.7 Ferramentas usadas para construção do site

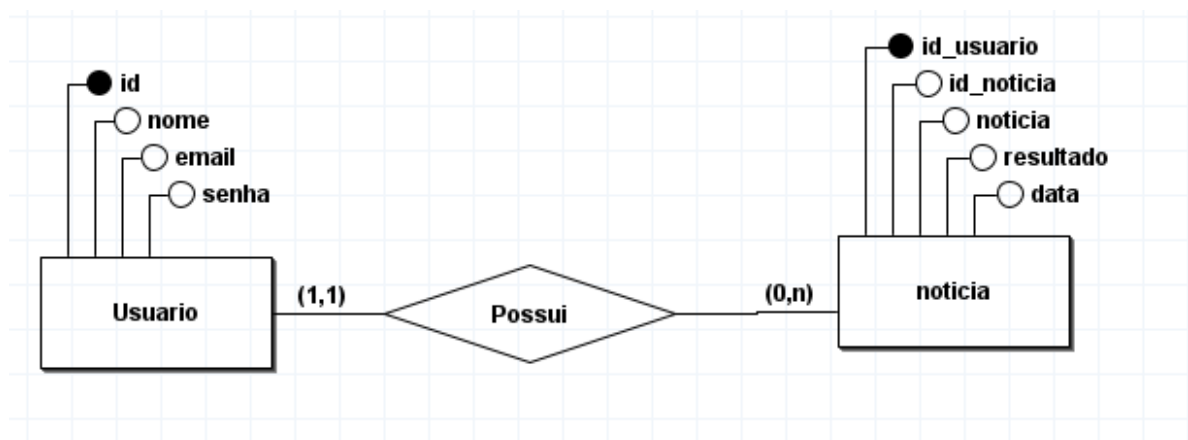
Para construção do site foi usada a arquitetura *MVC*, banco de dados *SQL* e o *SGBD MySQL*, linguagem *Python* e seu *framework Flask*, *HTML*, *Bootstrap*, *CSS* e *Javascript*

7.8 Banco de dados e modelagem

Como banco de dados foi usado *MySQL*, com o banco de dados tem o objetivo de cadastrar o usuário e salvar as notícias que ele consultado, salvando a notícia inteira, resultado (falsa ou verdadeira) e a data da consulta.

7.8.1 Modelo conceitual:

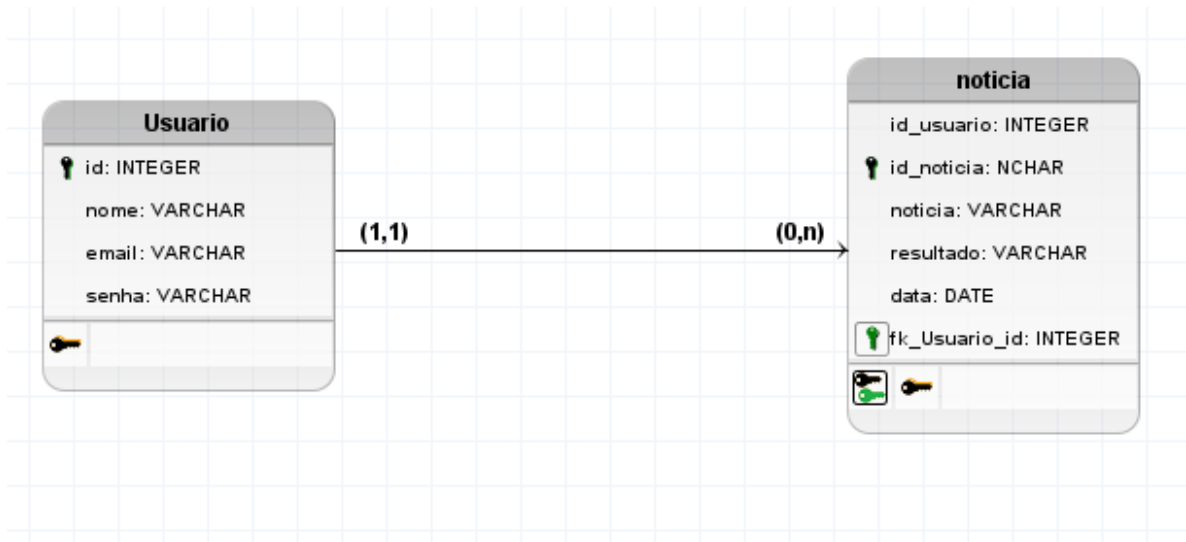
Figura 48 - Modelo conceitual do banco de dados



Fonte: Autores 2022

7.8.2 Modelo lógico

Figura 49 - Modelo lógico do banco de dados



Fonte: Autores 2022

7.8.3 Dicionário de dados:



7.9 Construção do site

Para construção do site utilizamos a arquitetura de software *MVC*, usando o *framework Flask*, e integrado com banco de dados *MySQL*. Nesse site é possível criar uma conta de usuário, alterar senha ou e-mail da sua conta, fazer a verificação da veracidade da notícia e consultar o histórico de notícias já verificadas.

Figura 50 - Tela de cadastro e login

FAKE NEWS

Detector

Cadastro

Nome:

E-mail:

Senha:

Reptla a senha:

Cadastrar

Login

E-mail:

Senha:

Logar

Esqueceu senha

Fonte: Autores 2022

Para fazer o cadastro, é necessário um nome, um e-mail e esse e-mail não poder ter um cadastro já com ele, e a senha que deve conter pelo menos uma letra maiúscula, minúscula, número, e deve ter pelo menos 8 caracteres, e essa senha é necessário que seja repetida para o usuário tenha certa do que digitou, e antes da senha ser salva no banco de dados ela é criptografada.

Figura 51 - Tela que auxilia a digitação da senha (quando está faltando algum requisito)

The image shows two side-by-side mobile app screens. The left screen is titled 'Cadastro' (Registration) and has a blue header. It contains input fields for 'Nome' (Name), 'E-mail', 'Senha' (Password), and 'Repita a senha' (Repeat password). The 'Senha' field has a single character visible. Below the fields is a grey box with the text 'A senha deve conter os seguintes passos:' (The password must contain the following steps:). It lists four requirements, each with a red 'X' icon: 'pelo menos uma letra minúscula' (at least one lowercase letter), 'pelo menos uma letra maiúscula' (at least one uppercase letter), 'pelo menos um número' (at least one number), and 'Mínimo 8 characters'. At the bottom is a green 'Cadastrar' (Register) button. The right screen is titled 'Login' and has a blue header. It contains input fields for 'E-mail' and 'Senha'. Below the 'Senha' field are two green buttons: 'Logar' (Login) and 'Esqueceu senha' (Forgot password).

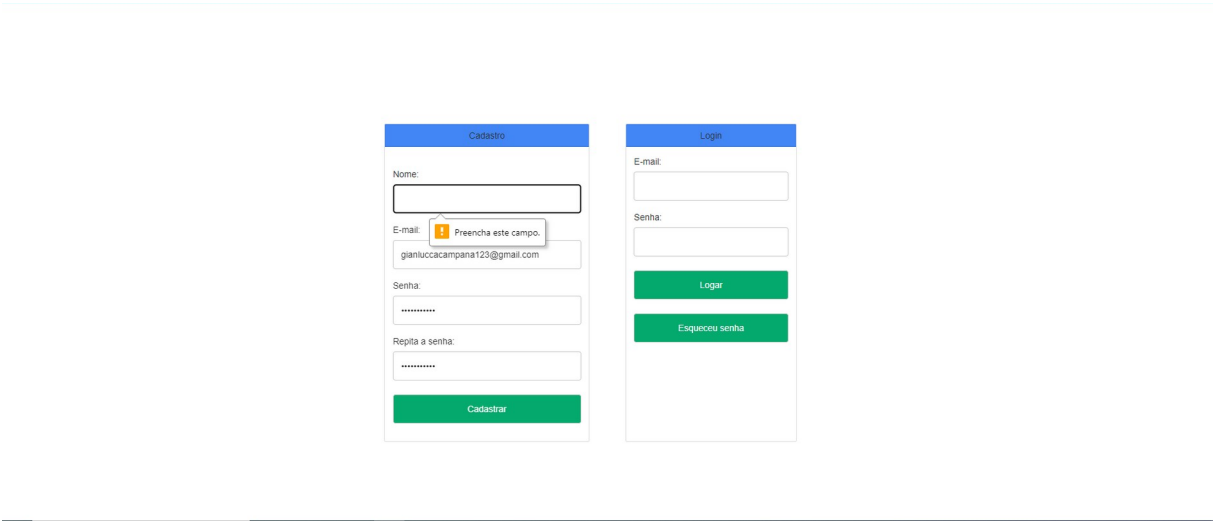
Fonte: Autores 2022

Figura 52 - Tela que auxilia a digitação da senha (quando todos requisitos estão corretos)

The image shows two side-by-side mobile app screens, similar to Figure 51. The left screen is titled 'Cadastro' (Registration) and has a blue header. It contains input fields for 'Nome' (Name), 'E-mail', 'Senha' (Password), and 'Repita a senha' (Repeat password). The 'Senha' field shows ten dots, indicating the password is masked. Below the fields is a grey box with the text 'A senha deve conter os seguintes passos:' (The password must contain the following steps:). It lists four requirements, each with a green checkmark icon: 'pelo menos uma letra minúscula' (at least one lowercase letter), 'pelo menos uma letra maiúscula' (at least one uppercase letter), 'pelo menos um número' (at least one number), and 'Mínimo 8 characters'. At the bottom is a green 'Cadastrar' (Register) button. The right screen is titled 'Login' and has a blue header. It contains input fields for 'E-mail' and 'Senha'. Below the 'Senha' field are two green buttons: 'Logar' (Login) and 'Esqueceu senha' (Forgot password).

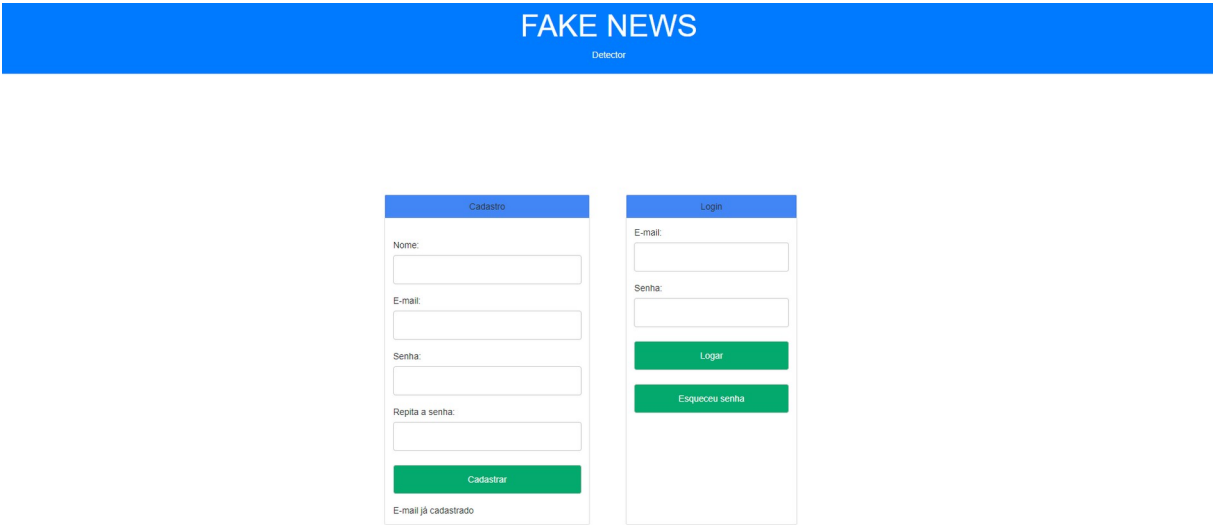
Fonte: Autores 2022

Figura 53 - Tela de cadastro quando falta preencher algum campo para o cadastro



Fonte: Autores 2022

Figura 54 - Tela de cadastro quando e-mail já foi cadastrado



Fonte: Autores 2022

Figura 55 - Tela cadastro quando as senhas digitadas são diferentes

FAKE NEWS

Detector

Cadastro

Nome:

E-mail:

Senha:

Repita a senha:

Cadastrar

Senha diferentes digitadas

Login

E-mail:

Senha:

Logar

Esqueceu senha

Fonte: Autores 2022

Figura 56 - Tela cadastro quando a senha não atende os requisitos mínimos

FAKE NEWS

Detector

Cadastro

Nome:

E-mail:

Senha:

Repita a senha:

Cadastrar

Senha não atende aos requisitos mínimos

Login

E-mail:

Senha:

Logar

Esqueceu senha

Fonte: Autores 2022

Figura 57 - Tela cadastro quando a conta é registrada

FAKE NEWS

Detector

Cadastro

Nome:

E-mail:

Senha:

Repita a senha:

Cadastrar

Conta registrada

Login

E-mail:

Senha:

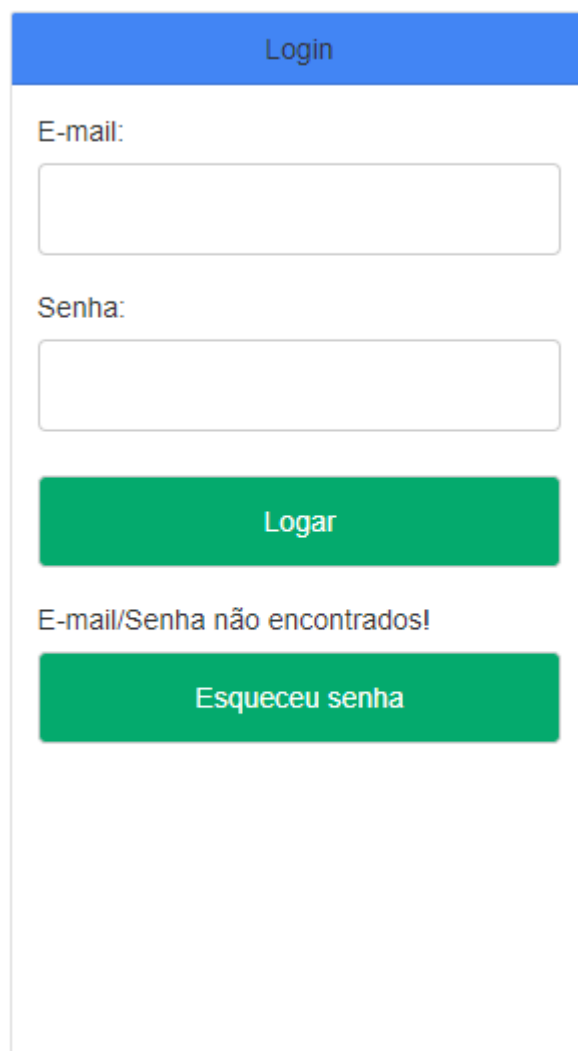
Logar

Esqueceu senha

Fonte: Autores 2022

Para efetuar login é necessário que já tenha o cadastro efetuado, e quando o login é efetuada, é aberta uma sessão e essa sessão possibilita ter acesso as outras páginas do site.

Figura 58 - Tela de login quando a sua senha ou e-mail não são encontrados no banco de dados



A interface de login apresenta um cabeçalho azul com o texto "Login". Abaixo, há campos para "E-mail:" e "Senha:", cada um com um input de texto. Um botão verde "Logar" está posicionado entre os campos. Abaixo do botão, uma mensagem de erro em vermelho indica "E-mail/Senha não encontrados!". Na base, há um botão verde "Esqueceu senha".

Login

E-mail:

Senha:

Logar

E-mail/Senha não encontrados!

Esqueceu senha

Fonte: Autores 2022

Efetutando o login já direcionado para página inicial do site, e nessa página é possível fazer a veracidade da notícia, no menu superior onde está o nome de usuário tem as opções de alterar e-mail, senha e deletar a conta, ver o histórico de notícias verificadas e sair da conta.

Figura 59 - Página inicial

FAKE NEWS

Detector

[HOME](#) [gianluca](#) * [Historico](#) [Sair](#)

Detector Noticia

Instruções

Para detectar a noticia corretamente, a noticia deve estar em **língua portuguesa** e conter no mínimo **100 palavras**.

Algoritmos Utilizados

Para realizar a detecção nos utilizamos de 3 algoritmos, Regressão Logística, SVC e MLP.

Enviar

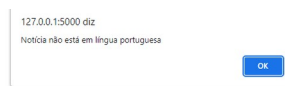
Criado por:

Mathues barnabe, Guilherme Vieira, Gianluca Campana

Fonte: Autores 2022

Ao colocar a notícia na caixa de texto e clicar em analisar, primeiro irá verificar se a notícia está em língua portuguesa, caso esteja em outra língua aparecerá uma mensagem informando o erro. Próxima verificação é se a notícia tem pelo menos 100 palavras, caso tenha de 100 palavras aparecerá uma mensagem falando da inconformidade. Passando das verificações a notícia é analisada e é dados seu resultado se é falsa ou não, e é salvo no banco de dados a notícia, resultado e data da verificação.

Figura 60 - Tela de análise quando a notícia não está língua portuguesa



Fonte: Autores 2022

Figura 61 - Tela de análise quando a notícia tem menos de 100 palavras



Fonte: Autores 2022

Figura 62 - Tela de análise quando a notícia tem o resultado de verdadeira:

[HOME](#) [galeria](#) [Histórico](#) [Sair](#)

Detector Noticia

Instruções

Para detectar a notícia corretamente, a notícia deve estar em **língua portuguesa** e conter no mínimo **100 palavras**.

Algoritmos Utilizados

Para realizar a detecção nos utilizamos de 3 algoritmos, Regressão Logística, SVC e MLP.

Enviar

Nossos algoritmos indentificaram a noticia como verdadeira, faça uma pesquisa antes de á divulgar! 😊

Fonte: Autores 2022

Figura 63 - Tela de análise quando a notícia tem o resultado de falsa

[HOME](#) [galeria](#) [Histórico](#) [Sair](#)

Detector Noticia

Instruções

Para detectar a notícia corretamente, a notícia deve estar em **língua portuguesa** e conter no mínimo **100 palavras**.

Algoritmos Utilizados

Para realizar a detecção nos utilizamos de 3 algoritmos, Regressão Logística, SVC e MLP.

Enviar

Nossos algoritmos indentificaram a noticia como uma possivel fakenews, faça uma pesquisa antes de á divulgar! 😞

Fonte: Autores 2022

Na página de alteração de e-mail (é necessário estar logado), é necessário que digite seu e-mail atual cadastrado, e o sistema irá verificar a sessão de e-mail criado no momento de login e igual ao e-mail digitado, e no campo abaixo irá digitar o novo e-mail desejado.

Figura 64 - Tela do site com e-mail não compatível com a sessão e-mail criada

A interface de alteração de e-mail apresenta um cabeçalho azul com o texto "Alterar Email". Abaixo, há dois campos de entrada: "Digite o e-mail atual:" e "Digite o e-mail novo:". Um botão verde com o texto "Alterar e-mail" está posicionado entre os campos. Na base da interface, uma mensagem de erro em grande e negrito afirma: "E-mail não correspondente".

Fonte: Autores 2022

Figura 65 - Tela do site com e-mail alterado

Alterar Email

Digite o e-mail atual:

Digite o e-mail novo:

Alterar e-mail

E-mail alterado com sucesso

Fonte: Autores 2022

Na página de alteração de senha (é necessário estar logado), é necessário que digite seu e-mail atual cadastrado e a senha atual cadastrada, e o sistema irá verificar se a sessão de e-mail e senha criadas no momento de login são iguais ao e-mail e senha digitada, e no campo abaixo irá digitar a nova senha desejada, que deverá conter pelo menos 8 caracteres e pelo menos uma letra maiúscula, uma minúscula e um número.

Figura 66 - Tela do site com e-mail não compatível com a sessão e-mail criada



The image shows a web form for changing a password. At the top is a blue header bar with the text "Alterar a senha". Below this are three input fields, each preceded by a label: "Digite o e-mail atual:", "Digite a senha atual:", and "Digite a senha nova:". Each input field is empty. Below the input fields is a green button with the text "Alterar e-mail". At the bottom of the form, there is a large, bold error message: "Senha ou E-mail não correspondente".

Alterar a senha

Digite o e-mail atual:

Digite a senha atual:

Digite a senha nova:

Alterar e-mail

Senha ou E-mail não correspondente

Fonte: Autores 2022

Figura 67 - Tela do site com senha não compatível com a sessão e-mail criada

Alterar a senha

Digite o e-mail atual:

Digite a senha atual:

Digite a senha nova:

Alterar e-mail

Senha ou E-mail não correspondente

Fonte: Autores 2022

Figura 68 - Tela do site com senha não a atendo aos requisitos mínimos

Alterar a senha

Digite o e-mail atual:

Digite a senha atual:

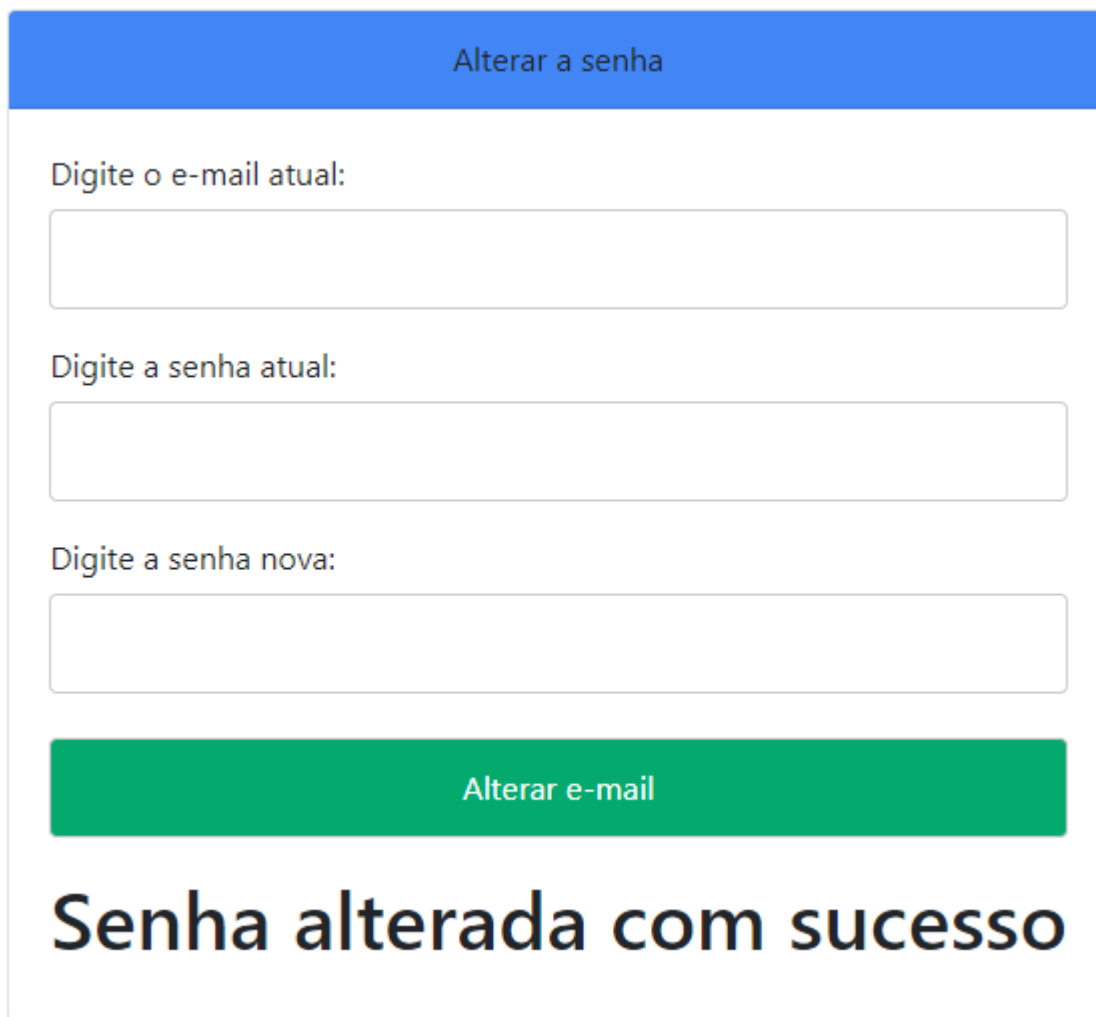
Digite a senha nova:

Alterar e-mail

Senha não atende os requisitos mínimos

Fonte: Autores 2022

Figura 69 - Tela do site com senha alterada com sucesso



The image shows a web form titled "Alterar a senha" (Change password) in a blue header. Below the header, there are three input fields with labels: "Digite o e-mail atual:" (Enter current email), "Digite a senha atual:" (Enter current password), and "Digite a senha nova:" (Enter new password). Each label is followed by an empty text input box. Below these fields is a green button labeled "Alterar e-mail" (Change email). At the bottom of the form, the text "Senha alterada com sucesso" (Password changed successfully) is displayed in a large, bold, black font.

Fonte: Autores 2022

Na página de deletar conta (é necessário estar logado), é necessário que digite seu e-mail atual cadastrado e a senha atual cadastrada, e o sistema irá verificar se a sessão de e-mail e senha criadas no momento de login são iguais ao e-mail e senha digitada, elas sendo compatíveis, a conta é deletada e volta para tela de cadastro:

Figura 70 - Tela do site com e-mail não compatível com a sessão e-mail criada

The screenshot shows a login interface with a blue header bar containing the text "Deletar". Below the header, there are two input fields: "Digite o e-mail:" and "Digite a senha:". Underneath these fields are two green buttons labeled "deletar" and "Voltar". At the bottom of the form, a large black error message reads "Senha ou E-mail não correspondente".

Fonte: Autores 2022

Figura 71 - Tela do site com senha não compatível com a sessão senha criada

This screenshot is identical to the one in Figure 70, showing the same login interface with the error message "Senha ou E-mail não correspondente".

Fonte: Autores 2022

Figura 72 - Tela do site com conta deletada

The image shows a web interface for 'FAKE NEWS Detector'. At the top, there is a blue header with the text 'FAKE NEWS' and 'Detector' below it. Below the header, there are two side-by-side forms. The left form is titled 'Cadastro' and contains four input fields: 'Nome:', 'E-mail:', 'Senha:', and 'Repita a senha:'. Below these fields is a green button labeled 'Cadastrar'. Below the button, it says 'Conta deletada com sucesso'. The right form is titled 'Login' and contains two input fields: 'E-mail:' and 'Senha:'. Below these fields is a green button labeled 'Logar'. Below the button, it says 'Conta deletada com sucesso' and a green button labeled 'Esqueceu senha'. At the bottom of the page, there is a blue footer with the text 'Criado por Mathues barnabe, Guilherme Vieira, Gianluca Campana'.

Fonte: Autores 2022

Na tela de histórico irá mostrar apenas as notícias que o usuário logado verificou, e será mostrado em uma tabela, que terá os 50 primeiros caracteres da notícia, o resultado dela com um sinal de *check* quando é verdadeira e um x quando falsa, e sua data de verificação e as notícias estarão ordenadas por data mais recente para a mais antiga. Para apresentar a notícia inteira é necessário clicar na notícia e mostrar ela inteira

Figura 73 - Tela histórico inicial

FAKE NEWS		
Detector		
HOME gianluca * Histórico Sair		
Noticias Testadas		
Notícia	Resultado	Data da Analise
Marcelo Miranda chegou ontem ao estado e fez uma r	x	2022-11-11 09:28:25
Mais de 90% das multas emitidas pela PRF (Policia	✓	2022-11-11 09:27:10
Mais de 90% das multas emitidas pela PRF (Policia	x	2022-11-11 09:26:50
Mais de 90% das multas emitidas pela PRF (Policia	✓	2022-11-11 09:24:21
primeira manha adriana ancelmo prisao domiciliar b	✓	2022-11-10 13:57:42
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:33:59
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:32:36
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:28:34
Janot pede ao STF 83 inquéritos para investigar po	✓	2022-11-01 11:23:49
O juiz Marcelo Bretas, da 7ª Vara Criminal Federal	x	2022-11-01 11:08:50
O juiz Marcelo Bretas, da 7ª Vara Criminal Federal	x	2022-10-26 10:27:09
Janot pede ao STF 83 inquéritos para investigar po	✓	2022-10-26 10:12:51

Fonte: Autores 2022

Figura 74 - Tela histórico com a notícia completa

FAKE NEWS		
Detector		
HOME gianluca * Histórico Sair		
Noticias Testadas		
Notícia	Resultado	Data da Analise
<p>Marcelo Miranda chegou ontem ao estado e fez uma r</p> <p>Marcelo Miranda chegou ontem ao estado e fez uma reunião com o primeiro escalão do governo, mas durante a manhã deste sábado (24) a Secretaria de Comunicação informou que ele voltou à Brasília para cumprir "compromissos partidários". Ele ficou menos de 24 horas no Tocantins. Cláudia compareceu, sem o governador, a um evento de entrega de cestas básicas a seis municípios que estão em situação de emergência por causa do volume de chuva. São eles: Cristalândia, Dueré, Formoso do Araguaia, Lagoa da Confusão, Pium e Santa Rita do Tocantins. A Defesa Civil entregou também colchões e kits de higiene para</p>	x	2022-11-11 09:28:25
Mais de 90% das multas emitidas pela PRF (Policia	✓	2022-11-11 09:27:10
Mais de 90% das multas emitidas pela PRF (Policia	x	2022-11-11 09:26:50
Mais de 90% das multas emitidas pela PRF (Policia	✓	2022-11-11 09:24:21
primeira manha adriana ancelmo prisao domiciliar b	✓	2022-11-10 13:57:42
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:33:59
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:32:36
hoje no dia vinte e cinco de novembro será lançado	x	2022-11-05 00:28:34
Janot pede ao STF 83 inquéritos para investigar po	✓	2022-11-01 11:23:49

Fonte: Autores 2022

Para sair da sessão é necessário clicar no botão que estará em todas as páginas do site, clicando nele a sua sessão atual será deletada e voltará para tela de login/cadastro

Na página inicial tem o botão de esqueceu a senha. Nesta página é preciso colocar o e-mail cadastrado para que aplicação consiga enviar o e-mail de alteração de senha. Antes de enviar o e-mail irá verificar no banco de dados se esse e-mail digitado já foi cadastrado, caso não tenha sido encontrado esse e-mail, o site mostrará uma mensagem que o e-mail não é cadastrado. Se o e-mail estiver cadastrado será enviado um e-mail com um link da página de alteração de senha, e será aberta uma sessão de esqueceu senha e essa sessão que irá possibilitar acessar a essa página. Na página de esqueceu a senha, a senha terá que ter pelo menos 8 caracteres, no mínimo uma letra maiúscula, uma minúscula e um número, e no campo a seguir repetir a senha digitada. Estando tudo certo a senha é alterada.

Figura 75 - Tela esqueceu senha com e-mail não encontrado

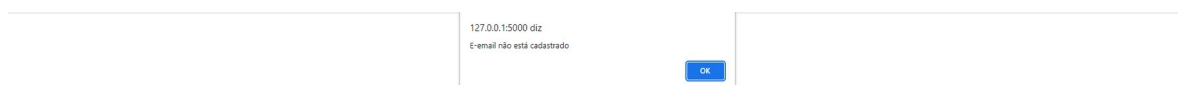
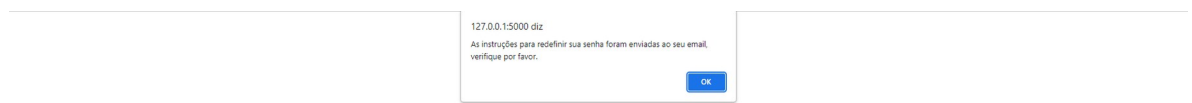


Figura 76 - Mensagem Enviada para o e-mail



Fonte: Autores 2022

Figura 77 - Tela mudar a senha com senhas digitas diferentes

FAKE NEWS

Detector

Mudar senha

Senha:

Repita a senha:

Atualizar senha

Senhas digitadas são diferentes

Criado por :

Mathues barnabe; Guilherme Vieira; Gianluca Campanna.

Fonte: Autores 2022

Figura 78 - Tela mudar senha com senha não atendendo requisitos mínimos

FAKE NEWS

Detector

Mudar senha

Senha:

Repita a senha:

Atualizar senha

Senha não atende aos requisitos mínimos

Criado por:

Mathues Barnabe; Guilherme Vieira; Gianluca Campanna

Fonte: Autores 2022

Figura 79 - Tela mudar senha alterada com sucesso

FAKE NEWS

Detector

Cadastro

Nome:

E-mail:

Senha:

Repita a senha:

Cadastrar

Login

E-mail:

Senha:

Logar

Esqueceu senha

Fonte: Autores 2022

Figura 80 - Tela mudar senha sem sessão



Fonte: Autores 2022

8 Considerações Finais

A aplicação feita ao longo do projeto foi implementada com banco de dados SQL, pelo relacionamento das tabelas e como SGDB foi utilizado o MySQL, por ser open source, muito utilizado no mercado de trabalho e é um software leve. Assim foi possível criar a tabela usuário e de notícias sendo relacionadas pelo id do usuário.

Para fazer o aprendizado de máquina foi utilizado o dataset com 7.200 notícias com 3.600 verdadeiras e 3.600 falsas, e utilizado modelo supervisionado, e para fazer o treinamento e o pré-processamento foi utilizado a linguagem Python na plataforma *collab* da Google. No pré-processamento foram retiradas pontuações, *stopwords*, aplicado *stemitização* e por último aplicado as técnicas *Bag of Words* e *TF-IDF*. Após o pré-processamento, foi feito o aprendizado de máquina e utilizado os algoritmos de predição SVM, Regressão Logística, MLPClassifier e Naives Bayes Multinomial e Bernoulli.

Com os resultados obtidos foram salvos em um arquivo *pickle* que foram o SVM, Regressão Logística e MLPClassifier e todos com a técnica TF-IDF e esse arquivo *pickle* é utilizado na página web para predição

Na criação da página web foi utilizado a linguagem Python e seu *framework* *Flask*, foram escolhidos o *Flask*, para manter o projeto com apenas uma linguagem e também porque ele é utilizado para projetos menores escalas diferentemente do *Django*.

Para a predição das notícias são utilizados os algoritmos que apresentaram os melhores resultados na predição, e fazer uma comparação de resultado entre eles, se maioria dos resultados forem como verdadeira, a notícia será salva como verdadeira e a mesma lógica é feita para as notícias falsas.

Com o término do projeto foi alcançado o objetivo proposto, que foi a criação de uma página web que fosse possível o usuário criar uma, fazer a verificação da veracidade de notícias e consultar todas as notícias que ele já verificou

Após o objetivo alcançado, observar que com o uso da inteligência artificial, aprendizado de máquina e seus algoritmos preditivos é possível minimizar um dos grandes problemas que a nossa sociedade enfrenta diariamente, que são as *Fake News* através do seu recebimento e disseminação. Não é à toa que os métodos utilizados e abordados ao longo deste projeto são utilizados nas áreas da saúde, financeiro, marketing, entre outras áreas que fazem a utilização de inteligência artificial e principalmente algoritmos de predição.

A oportunidade que tivemos para utilizar diferentes algoritmos de predição e diferentes técnicas de mineração de dados nos possibilitou abrir um leque com diferentes opções na qual nos sujeita a escolher o algoritmo que queremos que seja utilizado para a predição dos resultados, vantagens que nos trazem diversificação, variedade e confiança para com a acurácia das análises de notícias falsas.

E ainda é possível melhorar o trabalho já concluído, testando-o com outros algoritmos preditivos e outras técnicas de mineração de dados, e principalmente o aumento de dados em seu dataset, que quanto maior for a base de dados para o aprendizado de máquina, melhor será seu desempenho e a diminuição de dar um algum resultado com falso positivo ou falso negativo.

Referências

CARRIÇO, Enrico Soares; PIRES, Enzo Ulisses Maria; TERRA, Gabriel Campos Gomes; BASILIO, Matheus Ferraz Rocha. Impactos das fake news na sociedade e suas consequências jurídicas. **Jornal Eletrônico**. Minas Gerais, p. 197-217. jan. 2019. Disponível em: <https://www.jornaleletronicofivj.com.br/jefvj/article/view/795>. Acesso em: 10 nov. 2022.

AFP (Brasil). **Após fake news sobre cura de covid-19, 27 morrem no Irã por ingestão de álcool adulterado**. 2020. Disponível em: <https://noticias.uol.com.br/saude/ultimas-noticias/afp/2020/03/09/vinte-e-sete-pessoas-morrem-no-ira-depois-de-beber-alcool-adulterado-para-curar-covid-19.htm>. Acesso em: 09 mar. 2020.

PARANÁ. 2ª VICE PRESIDÊNCIA. **O perigo das fake news**. 2020. Disponível em: https://www.tjpr.jus.br/noticias-2-vice/-/asset_publisher/sTrhoYRKnlQe/content/o-perigo-das-fake-news/14797?inheritRedirect=false.%20Acesso%20em:%2001%20nov.%202022. Acesso em: 10 nov. 2022.

ORACLE (Brasil). **O que é Deep Learning?** Disponível em: <https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-deep-learning/>. Acesso em: 01 nov. 2022.

MDN WEB DOCS (org.). **JavaScript**. Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/JavaScript/About_JavaScript. Acesso em: 01 nov. 2022.

PISA, Pedro. **O que é e como usar o MySQL?** 2012. Disponível em: <https://www.techtudo.com.br/noticias/2012/04/o-que-e-e-como-usar-o-mysql.ghtml>. Acesso em: 02 nov. 2022.

NAIVE Bayes. Disponível em: https://scikit-learn.org/stable/modules/naive_bayes.html. Acesso em: 10 nov. 2022.

TAMAIAS, Ana Laura Moraes. **Modelos de Predição | Naive Bayes**. 2019. Disponível em: <https://medium.com/turing-talks/turing-talks-16-modelo-de-predicao-naive-bayes-6a3e744e7986#:~:text=Bernoulli%20Naive%20Bayes,valor%20dentre%20dois%20valores%20poss%C3%ADveis>. Acesso em: 02 nov. 2022.

QUIXADÁ, Carlos Matheus da Silva. **CLASSIFICAÇÃO DE TEXTO: MULTINOMIAL X BERNOULLI UTILIZANDO REVIEWS DE E-COMMERCE**. 2019. 38 f. Tese (Doutorado) - Curso de Engenharia de Software, Universidade Federal do Ceará, Ceará, 2019. Disponível em: https://repositorio.ufc.br/bitstream/riufc/44566/1/2019_tcc_cmsquixada.pdf. Acesso em: 02 nov. 2022.

CARRARO, Letícia Mendonça. **Modelos de Predição | Introdução à Predição**. Disponível em: <https://medium.com/turing-talks/turing-talks-10-introducao-naive-bayes-0-predicao-a75cd61c268d>. Acesso em: 02 nov. 2022.

MANNARA, Barbara. **44% dos brasileiros dizem receber fake news diariamente; veja pesquisa**. Disponível em: <https://www.techtudo.com.br/noticias/2022/08/44percent-dos-brasileiros-dizem-receber-fake-news-diariamente-veja-pesquisa.ghtml>. Acesso em: 02 nov. 2022.

COELHO, Alexandre Ramos. **Stemming para a língua portuguesa: estudo, análise e melhoria do algoritmo RSLP**. Disponível em: <https://lume.ufrgs.br/handle/10183/23576>. Acesso em: 02 nov. 2022.

RODRIGUES, Jéssica. **O que é o Processamento de Linguagem Natural?** Disponível em: <https://medium.com/botsbrasil/o-que-e-o-processamento-de-linguagem-natural-49ece9371cff>. Acesso em: 02 nov. 2022.

TIBCO. **O que é regressão logística?** Disponível em: <https://www.tibco.com/pt-br/reference-center/what-is-logistic-regression>. Acesso em: 02 nov. 2022.

DOCS, Mdn Web. **HTML básico**. Disponível em: https://developer.mozilla.org/pt-BR/docs/Learn/Getting_started_with_the_web/HTML_basics. Acesso em: 02 nov. 2022.

COUTINHO, Bernardo. **Modelos de Predição | SVM**. Disponível em: <https://medium.com/turing-talks/turing-talks-12-classifica%C3%A7%C3%A3o-por-svm-f4598094a3f1>. Acesso em: 02 nov. 2022.

ADDAN, Diego. **Support Vector Machine**. 2019. 33 f. TCC (Graduação) - Curso de Inteligência Artificial, Unibrasil, Paraná, 2019.

ADDAN, Diego. **Support Vector Machine**. Paraná: Diego Addan, 2019. 34 slides, color. Disponível em: <https://www.inf.ufpr.br/dagoncalves/IA07.pdf>. Acesso em: 02 nov. 2022.

STEINBRUCH, David. **Um estudo de algoritmos para classificação automática de textos utilizando naive-Bayes**. 2006. 75 f. Dissertação (Mestrado) - Curso de Informática, Puc Rio, Rio de Janeiro, 2006.

ORACLE (org.). **O que é um banco de dados relacional (RDBMS)?** Disponível em: <https://www.oracle.com/br/database/what-is-a-relational-database/>. Acesso em: 09 nov. 2022.

STEIN, Roger Alan; SILVA, Allan de Barcelos. Análise assintótica de algoritmo para geração de matriz termo-documento contendo TF-IDF. Figshare, [S.L.], v. 1, n. 9, p. 1-12, jan. 2016. Figshare. <http://dx.doi.org/10.6084/M9.FIGSHARE.4220691.V1>. Disponível em: https://www.researchgate.net/profile/Allan-Silva-5/publication/309920952_Analise_

assintotica_de_algoritmo_para_geracao_de_matriz_termo-
documento_contendo_TF-IDF/links/59e881d70f7e9bc89b540c0f/Analise-
assintotica-de-algoritmo-para-geracao-de-matriz-termo-documento-contendo-TF-
IDF.pdf. Acesso em: 10 nov. 2022.

X SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 2014, Londrina. Comparação entre TF-IDF e LSI para pesagem de termos em micro-blog. Londrina: Researchgate, 2014. Disponível em: https://www.researchgate.net/profile/Sylvio-Barbon-Junior/publication/262804841_Comparacao_entre_TF-IDF_e_LSI_para_pesagem_de_termos_em_micro-blog/links/0a85e53b316039746a000000/Comparacao-entre-TF-IDF-e-LSI-para-pesagem-de-termos-em-micro-blog.pdf. Acesso em: 10 nov. 2022.

A ORIGEM do CSS, um pouco da história. DEVMEDIA. Disponível em: <https://www.devmedia.com.br/a-origem-do-css-um-pouco-da-historia/15195>. Acesso em: 10 nov. 2022.

LIMA, Guilherme. Bootstrap: **O que é, Documentação, como e quando usar**. 2022. Disponível em: <https://www.alura.com.br/artigos/bootstrap>. Acesso em: 10 nov. 2022.