Sistema de Classificação de E-mails para Identificação de Spam utilizando o Classificador Ingênuo de Bayes

1st Amanda Cristina Fernandes M. de Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
acfml@cin.ufpe.br

2nd Maria Letícia do N. Gaspar
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mlng@cin.ufpe.br

3rd Mariana Melo dos Santos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mms11@cin.ufpe.br

4th Victória Barbosa Cesar Figueiredo
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
vbcf@cin.ufpe.br

***Abstract*—Este projeto visa utilizar o classificador probabilístico Ingênuo de Bayes para identificação eficiente de Spam, distinguindo mensagens indesejadas de conteúdos legítimos através da análise de atributos textuais e estruturais.**

***Index Terms*—email, identificação de spam, classificação, Naïve Bayes**

I. Introdução

Spam é uma prática de enviar e-mail ou mensagens sem o consentimento do usuário na internet. Essas mensagens possuem conteúdos indesejados ou perigosos para o usuário como: Propagandas, vírus, fake news ou estelionato. Segundo o SpamLaws, site especializado em proteção na internet, 45% de todos os e-mails enviados por dia são spam. Além disso, 20% dos e-mails, principalmente de marketing, serão classificados como spam incorretamente e nunca chegarão a potenciais consumidores. Assim, é preciso encontrar uma forma de detectar se os e-mails são ou não spam para que o usuário não receba diversas mensagens desagradáveis ou perigosas e também não deixe de receber e-mails de seu interesse.

II. Objetivos

O principal objetivo deste projeto é desenvolver um modelo eficiente capaz de calcular de maneira precisa e confiável a probabilidade de um e-mail ser classificado como spam ou não, a partir da construção de um filtro personalizado que considera aspectos da estrutura e do conteúdo da mensagem, além da avaliação dos usuários. Para isso, será utilizado a base de dados 'Spambase'.

III. Justificativa

A utilização de e-mails continua sendo uma das principais formas de comunicação digital e é amplamente utilizada em diversos contextos, tanto pessoais quanto profissionais. No entanto, todos os dias, os e-mails de spam tornam-se mais inteligentes e mais sábios, apresentando um conteúdo melhor e desenvolvendo novos métodos capazes de ultrapassar os filtros de spam empregados atualmente.

Nesse contexto, existe a necessidade de desenvolver um modelo de filtragem baseado no algoritmo *Naïve Bayes* tanto para evitar que e-mails legítimos sejam classificados erroneamente como spam, quanto para que a caixa de entrada de usuários não seja poluída de conteúdos indesejados e possivelmente perigosos.

IV. Metodologia

Neste projeto, buscaremos desenvolver um modelo baseado em machine learning capaz de diferenciar e-mails legítimos de spams. Para alcançar esse objetivo, faremos uma análise exploratória dos dados para entender o comportamento e identificar os principais padrões textuais e estruturais característicos de e-mails ilegítimos. Para esse fim, iremos utilizar, principalmente, o algoritmo classificador *Naïve Bayes* da biblioteca *Scikit-learn*.

Todo o código será desenvolvido em *Python*, em um ambiente de desenvolvimento do *Google Collaboratory*. Utilizaremos as principais bibliotecas para análise e modelagem de dados, visualização gráfica e operações matemáticas construídas para o *Python: Scipy, Numpy, Pandas, Matplotlib, Seaborn e Scikit learn*.

A. Dataset

O conjunto de dados escolhido para análise foi a *Spambase*, disponibilizado pelo Repositório de Aprendizado de Máquina UC Irvine (*UC Irvine Machine Learning Repository*) . Esse se utiliza de algoritmos de aprendizagem de máquina para

classificação binária de amostras de email em spam ou não-spam. Nessa abordagem, escolhemos a Modelo Ingênuo de Bayes como referência para analisar esse conjunto de dados e, assim, desenvolver um novo modelo de Classificação de Spam.
A maioria dos atributos presentes nessa base de dados indica se uma determinada palavra ou caractere estava ocorrendo com frequência no e-mail. Aqui seguem as definições desses atributos:

1) 48 atributos reais contínuos [0, 100] do tipo word_freq_WORD:
Porcentagem de palavras no e-mail que correspondem ao WORD, ou seja, 100 * (número de vezes que o WORD aparece no e-mail) / número total de palavras no e-mail. Uma ”palavra”, nesse caso, é qualquer cadeia de caracteres alfanuméricos delimitada por caracteres não alfanuméricos ou fim de cadeia de caracteres.

2) 6 atributos reais contínuos [0, 100] do tipo char_freq_CHAR:
Porcentagem de caracteres no e-mail que correspondem a CHAR, ou seja, 100 * (número de ocorrências CHAR) / total de caracteres no e-mail.

3) 1 atributo real contínuo [1,...] do tipo capital_run_length_average:
Comprimento médio das sequências ininterruptas de letras maiúsculas.

4) 1 atributo inteiro contínuo [1,...] do tipo capital_run_length_longest:
Comprimento da maior sequência ininterrupta de letras maiúsculas.

5) 1 atributo inteiro contínuo [1,...] do tipo capital_run_length_total:
Soma do comprimento das sequências ininterruptas de letras maiúsculas.

6) 1 atributo de classe nominal $\{0,1\}$ do tipo spam:
Indica se o e-mail foi considerado spam (1) ou não (0), ou seja, e-mail comercial não solicitado.

B. Processamento do dataset

O processamento que iremos aplicar no *dataset* consiste em 3 etapas principais: (1) Filtro de Instâncias, (2) Seleção de Atributos e (3) Engenharia de Atributos:

- Filtro de Instâncias: Nesta fase, procederemos à correção ou remoção de valores ausentes, considerando a abundância de informações incompletas no banco de dados. Além disso, eliminaremos quaisquer instâncias duplicadas, redundantes ou irrelevantes para a análise em questão.

- Seleção de Atributos: Nem todos os atributos presentes no conjunto de dados serão considerados como entradas para os modelos de aprendizado. Esta fase consiste em escolher um conjunto de atributos mais adequado para o modelo em questão. Tal conjunto deve incluir somente os atributos mais relevantes para a análise, além de apresentar dimensão apropriada.

- Engenharia de Atributos: Nesta etapa acontecem transformações dos atributos a fim de permiter que esses sejam utilizados pelos algoritmos de aprendizado e, até mesmo, melhorar o desempenho dos modelos. Caso necessário, serão aplicadas técnicas de *encoding* para transformar dados categóricos em dados numéricos, de modo que possam ser utilizados pelos algoritmos de aprendizado.

C. Algoritmos de aprendizagem de máquina

Aprendizagem de máquina é uma área da inteligência artificial que permite o aprimoramento de computadores através de bancos de dados.Nesse contexto a máquina irá reconhecer os padrões existentes. O aprendizado de máquina pode ser classificado em quatro tipos: Aprendizado supervisionado, aprendizado não supervisionado, semi-supervisionado e aprendizado por reforço.
Nesse projeto,utilizaremos o algoritmo ingênuo de Bayes que é um algoritmo supervisionado. Esse classificador foi criado por Thomas Bayes que nasceu em Londres e foi um pastor presbiteriano e matemático. Além disso, foi eleito membro da Royal Society em 1742.
O algoritmo de Bayes é aplicado, por exemplo, em classificadores de sentimentos e na área de saúde para determinar se há ou não a ocorrência de uma doença. Só não é muito indicado para casos muito complexos que envolvam uma quantidade muito grande de dados ou variáveis numéricas.

Esse modelo considera as features (variáveis) independentes entre si e possui uma base matemática relativamente simples e também um bom desempenho em comparação com os demais classificadores.

Uma vantagem desse algoritmo é que ele depende de poucas dados para ter uma boa acurácia. O algoritmo funciona da seguinte maneira: para calcular a predição, o algoritmo irá definir uma tabela de probabilidades, em que consta a frequência dos preditores com relação às variáveis de saída. Então, o cálculo final leva em conta a probabilidade máxima para oferecer uma solução.
O algoritmo utiliza o Teorema de Bayes (cálculo da probabilidade de um evento dado que outro evento já ocorreu) que possui a fórmula:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

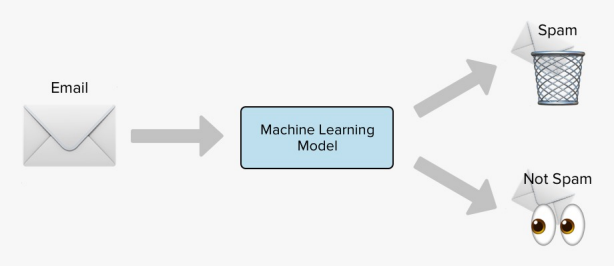

Fig. 1. Ilustração do funcionamento do filtro de spam

- P(A) : Probabilidade do evento A ocorrer
- P(B) : Probabilidade do evento B ocorrer.
- $P(B|A)$: Probabilidade de B ocorrer sabendo que A já ocorreu.
- $P(A|B)$: Probabilidade de A ocorrer sabendo que B já ocorreu.

Considerando o evento A como a ocorrência de spam e o evento B como a ocorrência de uma determinada palavra, teremos a aplicação desse conceito na classificação de e-mails como spam e não spam.
A partir dessa métrica, é possível desenvolver um filtro capaz de detectar os e-mails indesejados e separá-los dos legítimos. Isso é feito por meio de uma medição do quanto de lixo eletrônico cada e-mail está enviando, ou seja, é calculada a probabilidade de cada palavra no e-mail ser spam e, então, tais probabilidades são multiplicadas pela quantidades de ocorrências de cada termo no e-mail, obtendo-se a métrica geral de spam capaz de classificar tais mensagens.

REFERENCES

[1] M. Paul, "Probabilidade: Aplicações à Estatítica". 2 Edição. livros Técnicos e Científicos Editora.
[2] Spambase [https://archive.ics.uci.edu/dataset/94/spambase]
[3] https://www.spamlaws.com/spam-stats.html
[4] https://kinsta.com/pt/blog/por-que-meus-e-mails-estao-indo-para-o-spam/
[5] https://www.one.com/pt/email/o-que-e-um-filtro-de-spam-e-virus
[6] https://www.bbc.com/portuguese/internacional-59701523
[7] https://www.voitto.com.br/blog/artigo/teorema-de-bayes

V. CRONOGRAMA DE ATIVIDADES

Apresentamos na Tabela I, o cronograma de atividades planejadas para o projeto final durante o período de entrega.

TABLE I
ATIVIDADES

Data	Atividades
18/07	***Seleção do Dataset***
20/07	***Elaboração de ideias***
24/07	***Escrita da proposta***
01/08	***Entrega da proposta***
07/08 - 13/08	***Desenvolvimento do projeto***
14/08	***Finalização do projeto***
16/08	***Elaboração da apresentação/ Elaboração dos slides***
18/08	***Gravação da apresentação***
21/08	***Possíveis correções***
31/08 - 12/09	***Entrega do projeto***

No cronograma, abordamos desde a concepção desta proposta até o desenvolvimento e entrega do projeto. A finalidade deste planejamento é estabelecer o tempo necessário para cada atividade que compõe o projeto final, bem como definir os marcos de progresso da equipe em relação às atividades e seus prazos limite (deadline).