16º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2025

Estudo Comparativo entre a Inteligência Artificial Generativa e o Aprendizado de Máquina com Processamento de Linguagem Natural Aplicado à Detecção de E-mail Spam

VITOR H. DOS SANTOS[1], ADRIANO J. FERRUZZI[2]

[1] Estudante do Curso Técnico Integrado em Redes de Computadores, Bolsista PIBIFSP, IFSP, Campus Pirituba, santos.hugo1@aluno.ifsp.edu.br
[2] Professor do Curso Técnico Integrado em Redes de Computadores, IFSP, Campus Pirituba, adriano.ferruzzi@ifsp.edu.br
Área de conhecimento (Tabela CNPq): 1.03.00.00-7 Ciência da Computação

RESUMO: E-mails são um dos principais meios de comunicação na internet. Através deles são propagados os spams que podem ser vistos como inofensivos, porém é um dos principais recursos na aplicação de golpes cibernéticos, golpes esses que envolvem o uso de engenharia social para enganar a vítima e se apossar de suas informações pessoais, muitas vezes de natureza financeira. Esse tipo de golpe é utilizado devido ao seu baixo custo e pouco conhecimento necessário por parte do golpista. Tendo em vista a importância de proteger as pessoas de golpes por meio de spam, o seguinte projeto visa estudar, desenvolver e aplicar modelos de processamento de linguagem natural (PLN) e aprendizado de máquina (ML) para detecção e classificação de e-mails spam e e-mails legítimos, em seguida, o trabalho usará a um grande modelo de linguagem para fazer a detecção das mensagens spam. Finalmente, a pesquisa fará a comparação entre os resultados obtidos pelas técnicas de classificação e apresentará qual técnica faz as classificações com maior assertividade.

PALAVRAS-CHAVE: processamento de linguagem natural; aprendizado de máquina, grande modelo de linguagem.

Comparative Study between Generative Artificial Intelligence and Machine Learning with Natural Language Processing Applied to E-mail Spam Detection

ABSTRACT: E-mails are one of the primary means of communication on the internet. They are also a major channel for the dissemination of spam, which, although sometimes perceived as harmless, represents a key vector in the execution of cyber scams. These scams often rely on social engineering techniques to deceive victims and gain access to their personal information, frequently of a financial nature. This method has been employed due to its low cost and the minimal technical knowledge required by the perpetrator. Due to the importance of protecting individuals from scams carried out through spam, this project aims to study, develop, and implement natural language processing (NLP) and machine learning (ML) models to detect and classify spam and legitimate e-mails. Subsequently, the study will employ a large language model to perform spam message detection. Finally, the research will compare the results obtained through the different classification techniques and identify which method yields the highest accuracy in classification.

KEYWORDS: natural language processing; machine learning, large language model.

INTRODUÇÃO

Com o avanço das técnicas de inteligência artificial (IA) e, mais recentemente, com o conhecimento popular das redes generativas, muitos golpistas têm se aproveitado para aprimorar seus métodos de ataque. No caso dos e-mails spam, seus textos podem ser corrigidos e aperfeiçoados, a fim de dificultar sua identificação e aumentar a sua credibilidade.

De acordo com OLIVO et al (2015), uma prática comum entre os golpistas que disseminam *phishing* utilizando de e-mails spam é valer-se de um texto persuasivo e convincente, que se assemelha muito com mensagens verdadeiras, mas, na verdade, o remetente visa enganar o usuário e convencê-lo a realizar alguma ação que possa torná-lo vítima de algum tipo de fraude.

Tem-se, portanto, a necessidade do estudo de técnicas de classificação as quais facilitam a identificação desses e-mails mal intencionados, como modelos de aprendizado de máquina (ML) com técnicas de processamento de linguagem natural (PLN) e também de redes neurais generativas que utilizam grandes modelos de linguagem.

Dessa forma, a presente pesquisa visa desenvolver e avaliar modelos de ML e PLN na classificação de e-mails spam, comparando seus resultados com a capacidade de classificação da inteligência artificial generativa brasileira MariTalk, um *large language model* (LLM) especializado na língua portuguesa. A pesquisa tem o intuito de demonstrar a capacidade de diferentes técnicas de classificação na detecção de spam, apresentando qual delas irá obter o melhor resultado.

MATERIAL E MÉTODOS

Este estudo está metodizado em pesquisas teóricas e aplicações práticas, sendo feito por meio de revisão e levantamento bibliográfico de natureza qualitativa. Ele está desenvolvido com base em artigos publicados nos últimos 10 anos, com exceção do dataset de treinamento que foi publicado em 2003, esse dataset foi escolhido devido a estrutura em que as mensagens estão organizadas. Para encontrar os artigos, foram utilizadas as bases de dados públicas disponíveis em sites como Periódicos Capes e Google Acadêmico.

A respeito da escolha dos artigos, foram utilizadas as seguintes palavras-chave no Google Acadêmico: PLN, e-mail spam, redes neurais generativas, aprendizado de máquina e engenharia social. A respeito dos critérios de inclusão, foram considerados os artigos que estejam nos idiomas português e inglês relacionados ao tema, que contenham conteúdo relevante e próximo ao da presente pesquisa, foram desconsiderados os que não estavam nos idiomas estabelecidos acima e não apresentem similaridades com a temática base desta pesquisa.

Além disso, foi realizado um levantamento de bases de dados públicas disponíveis, tendo como objetivo encontrar dados de e-mails pré-classificados em spam e legítimo. Esse banco de dados serve como base para o treinamento e testes dos modelos.

Para a criação do modelo, foi necessário realizar três principais etapas, sendo elas: análise exploratória, limpeza e transformação dos dados, e, por fim, o treinamento do modelo de ML, o qual, uma vez criado, é responsável pela classificação de e-mails spam. Em seguida, será utilizado um LLM para fazer as classificações dos e-mails.

O treinamento foi feito com o uso das variáveis “origem do e-mail”, “subject” e “mensagem” como base para fazer as classificações. O banco de dados ainda possui uma variável alvo que identifica se a mensagem é um e-mail legítimo ou um spam.

Para o armazenamento do banco de dados selecionado foi utilizado o Google Drive. O motivo da escolha dessa plataforma é a integração com o Jupiter Notebook, ambas gratuitas, que permitem a troca rápida de interfaces, a qual apresenta como *back-end* uma máquina virtual que opera como um nó computacional, possuindo 2 núcleos Intel Xeon de 2.20 GHz, 12 GB de memória RAM e 1 GPU Tesla T4 com 15 GB de memória interna. Além disso, ela dá suporte a bibliotecas de ML que são importantes para o desenvolvimento dos algoritmos.

FUNDAMENTAÇÃO TEÓRICA

A engenharia social pode ser definida como uma técnica de manipulação que explora erros humanos com o objetivo de obter suas informações pessoais, acessos não autorizados e coisas de valor. Esses ataques não se restringem apenas ao on-line e podem também acontecer em pessoa ou por outros meios de interação e comunicação (AVANCI et al, 2022).

O spam pode ser definido como o envio indiscriminado de e-mails sem o consentimento de seus destinatários. O envio de spam pode acarretar em diversos problemas, desde o incômodo a quem o recebe, até a sobrecarga de servidores SMTP (*Simple Mail Transfer Protocol*), por conta da grande quantidade de e-mails recebidos. Em situações mais graves, como no caso dos e-mails *phishing*, o dano causado pode ir muito além de apenas um incômodo, podendo comprometer a segurança dos dispositivos de usuários, bem como acarretar em perdas financeiras (OLIVO et al, 2015).

Atualmente, com a popularização da IA, golpistas se aproveitam para sofisticar seus métodos de ataque, principalmente os que envolvem engenharia social, como por exemplo, na elaboração de ataques que podem se propagar por spam*,* como *phishing* e o *ransomware*, que tem seus erros gramaticais e ortográficos corrigidos com o uso de IAs, trazendo maior credibilidade para o texto (CISO ADVISOR, 2024).

Para ajudar a fazer a classificação dos e-mails, pode-se usar modelos de ML. Trata-se de uma área da IA que permite computadores aprenderem e melhorarem seu desempenho em tarefas específicas a partir de um conjunto de dados, não necessitando de uma programação explícita. Esses sistemas utilizam uma grande quantidade de dados no intuito de identificar padrões e prever resultados (VALDATI, 2020).

Alguns algoritmos de classificação podem ser propostos para colaborar na avaliação de milhares de mensagens em pouco tempo. Esses algoritmos de ML podem encontrar padrões nas mensagens e identificar uma mensagem verdadeira de uma mensagem spam.

Segundo VALDATI (2020), técnicas de aprendizado de máquina devem ser designadas para tarefas que não possam ser resolvidas de maneira determinística, tendo a necessidade de extrair conhecimentos com base em exemplos. Elas podem ser classificadas em aprendizado supervisionado e não supervisionado.

No aprendizado supervisionado, o sistema é treinado com base em um conjunto de dados rotulados, ou seja, dados que já possuem uma resposta correta. No contexto de detecção de spam, todos os e-mails utilizados para o treinamento do modelo já vão estar pré-rotulados como “spam” ou “não spam”. No aprendizado não supervisionado o modelo é treinado com base em um conjunto de dados que não possuem rótulos. Nesse caso, o objetivo é identificar padrões e estruturas nos dados, agrupando os semelhantes.

Segundo FONTANA (2020), “Algoritmos de classificação são algoritmos de aprendizagem supervisionada onde o objetivo é prever uma classe ou rótulo associado com uma variável de entrada contendo determinados atributos”. Dessa forma, os algoritmos de classificação são ótimos aliados na tarefa de separar os dados em classes, como spam e não spam. Dentre os diversos algoritmos de classificação, cada um deles utiliza uma abordagem específica na classificação dos dados.

O algoritmo de Naive Bayes, segundo LEITE et al (2021), é um classificador probabilístico que utiliza uma abordagem de aprendizado supervisionado. Esse algoritmo é baseado no teorema de Bayes, que é utilizado na determinação dos resultados da classificação, em que assume não haver dependências entre as variáveis do sistema. Uma das principais vantagens apresentadas pelo modelo é a sua capacidade de classificar dados aos quais ele nunca foi treinado.

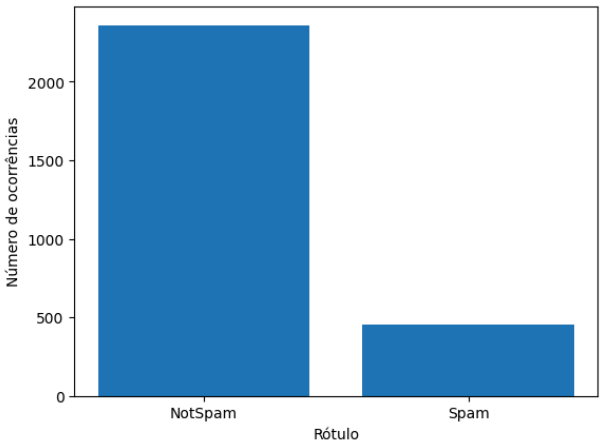
Já o SVM, é um método de aprendizado de máquina que pode ser utilizado em problemas de classificação e regressão, além de outras tarefas de aprendizagem. Os autores destacam a boa capacidade de generalização do algoritmo, que se baseia no conceito de planos de decisão, que definem seus limites construindo hiperplanos em um espaço multidimensional. Esses hiperplanos são responsáveis por separar os dados da amostra em categorias diferentes (LEITE et al, 2021).

Segundo SILVA et al (2023), as árvores de decisão são algoritmos de aprendizado de máquina supervisionado que se fundamentam em uma estrutura em forma de árvore para basear suas regras de decisão. O modelo de *Random Forest* é uma extensão da técnica das árvores de decisão, que utiliza um conjunto dessas árvores para suas previsões. Isso permite uma maior robustez e generalização por parte do modelo quando comparado às árvores de decisão individuais.

RESULTADOS E DISCUSSÃO

Os algoritmos de ML *Support Vector Machine* (*SVM*), *Naive Bayes* e *Random Forest* foram aplicados à uma base de dados pública contendo 2814 mensagens de e-mail, sendo 455 mensagens classificadas como spam e 2359 mensagens classificadas como e-mails legítimos (ou, *Not* Spam), estando todas elas em inglês. Sendo assim, a base de dados utilizada apresenta 16.17 % de mensagens rotuladas como positivas em relação aos e-mails indesejados (SAKKIS, 2003). A quantidade das amostras pode ser observada na Figura 1.

FIGURA 1. Distribuição de Rótulos da Base de Dados.



Fonte: Elaborado pelo autor (2025).

Seguidamente, apresenta-se uma tabela que contém os resultados dos testes dos respectivos modelos utilizados, sendo eles o *Naive Bayes, Random Forest* e *SVM*. Os dados foram baseados nos resultados da função *macro average*, que realiza uma média simples entre as performances de classificação dos modelos nas classes Spam e Não Spam.

TABELA 1. Resultados das análises dos modelos de aprendizado de máquina.

Teste	Naive Bayes	Random Forest	Support Vector Machine
Precisão	0.98	0.99	1.00
Revocação	0.98	0.94	0.98
F1-Score	0.98	0.96	0.99
Acurácia	0.99	0.98	0.99

Fonte: Elaborado pelo autor (2025).

Tendo em vista os dados demonstrados na tabela, fica evidente que o modelo *SVM* obteve a melhor performance, apresentando métricas como a de precisão (1.00) com nenhum falso positivo ou falso negativo. Além disso, apresentou também taxas altas em outros testes, como no F1-score e na acurácia geral do modelo (0.99), e na métrica de revocação (0.98). É importante salientar que foi necessário redimensionar os hiperparâmetros do modelo para evitar o *overfitting*, o que o ajudou significativamente na generalização, obtendo desempenhos ainda maiores nos testes.

O algoritmo de *Naive Bayes* conseguiu resultados equilibrados, como pode-se observar nas métricas de precisão, revocação e F1-score, em que todas registraram 0.98. Concluindo com uma acurácia excelente (0.99). Aqui, também devido ao *overfitting* observado, foi necessário realizar o redimensionamento dos hiperparâmetros do modelo, melhorando expressivamente seus resultados, o que também contribuiu para a estabilidade do algoritmo.

Por último, o *Random Forest*, que atingiu uma precisão ligeiramente superior (0.99), em comparação com o *Naive Bayes*, mas que por outro lado ficou atrás em comparação aos outros modelos nos testes de revocação e F1-score, apresentando, respectivamente, 0.94 e 0.96 como suas médias. Por sua vez, a acurácia geral (0.98), embora também ligeiramente inferior às dos outros algoritmos, pode ser considerada muito boa. Ainda mais se for considerado a não necessidade de redimensionar os hiperparâmetros do algoritmo de *Random Forest*, mesmo apresentando *overfitting*. Já que, segundo LEITE et al. (2021) por ser um algoritmo do tipo *ensemble learning*, ou seja, por agrupar os resultados ou as previsões de diversas árvores de decisão que foram treinadas individualmente, esse modelo diminui a variância e o viés de seus resultados.

CONCLUSÕES

Após analisar e comparar os resultados obtidos pelos modelos, pode-se concluir que todos se mostraram bons na classificação de e-mails spam, com valores altos nas métricas de precisão, recall, F1-score e acurácia. Entretanto, o modelo SVM demonstrou um melhor equilíbrio entre os testes, principalmente após ter seus hiperparâmetros redimensionados, a fim de evitar o *overfitting*. O algoritmo de Naive Bayes também se mostrou uma ótima opção, apresentando um desempenho equilibrado, além de uma ótima generalização após o redimensionamento de hiperparâmetros. Por fim, dado os resultados apresentados pelo modelo de *Random Forest*, posto que não foi necessário realizar ajustes, o algoritmo teve seu recall comprometido, sendo relevante quando a detecção de spam é a prioridade. Portanto, embora todos os modelos tenham se mostrado viáveis na tarefa de classificar e-mails spam, o SVM representou os melhores resultados para esta finalidade.

De agora em diante, pretende-se utilizar da API da Maritaca AI, uma IA generativa brasileira que é especializada na língua portuguesa e que disponibilizou gratuitamente créditos de sua API para esta pesquisa, com o intuito de classificar os e-mails da mesma base de dados já utilizada nos modelos de ML. A fim de demonstrar qual método consegue obter os melhores resultados na tarefa de classificação de e-mails spam.

CONTRIBUIÇÕES DOS AUTORES

O autor fez o levantamento da fundamentação teórica e da base de dados, realizou a parte prática da pesquisa e redigiu o trabalho.

Todos os autores contribuíram com a revisão do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Agradeço a Deus, por ter me dado forças para superar os desafios na execução da pesquisa. A minha família, pelo apoio e incentivo na realização do projeto. E ao meu orientador, pela atenção e colaboração no desenvolvimento do projeto. Os autores agradecem o IFSP Campus Pirituba pela disponibilização dos recursos para o projeto.

REFERÊNCIAS

AVANCI DE SOUZA PEREIRA, L.; VICENTINE, A. L.; CASTRO RIZO, A. **Impactos da Engenharia Social na Segurança da Informação**. *Revista Brasileira em Tecnologia da Informação*, [S. l.], v. 4, n. 1, p. 48–58, 2022. Disponível em: https://www.fateccampinas.com.br/rbti/index.php/fatec/article/view/75. Acesso em: 28 out. 2024.

CISO ADVISOR. **Cibercriminosos usam IA em táticas de ransomware**. *CISO Advisor*. Disponível em: https://www.cisoadvisor.com.br/branded-posts/cibercriminosos-usam-ia-em-taticas-de-ransomware/. Acesso em: 01 ago. 2024.

FONTANA, É. **Introdução aos Algoritmos de Aprendizagem Supervisionado**. Disponível em: https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf. Acesso em: 13 nov. 2024.

GALEANO, S. R. **Zero-Shot Spam Email Classification Using Pre-trained Large Language Models**. Maio de 2024. Disponível em: https://arxiv.org/abs/2405.15936. Acesso em: 15 Abril. 2025.

LEITE, Danilo Rangel Arruda; MORAES, Ronei Marcos de; LOPES, Leonardo Wanderley. Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando Random Forest. **Journal of Health Informatics**, Brasil, v. 12, 2021. Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/814. Acesso em: 22 jun. 2025.

OLIVO, C.; SANTIN E LUIZ, A.; OLIVEIRA, E. **Capítulo 4 Abordagens para Detecção de Spam de E-mail**. XV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais. Disponível em: https://secplab.ppgia.pucpr.br/files/papers/2015-7.pdf. Acesso em: 22 jun. 2025.

SAKKIS, Georgios; ANDROUTSOPOULOS, Ion; PALIOURAS, Georgios; KARKALETSIS, Vangelis; SPYROPOULOS, Constantine D.; STAMATOPOULOS, Panagiotis. **A Memory-Based Approach to Anti-Spam Filtering**. ResearchGate**,** 2003. Disponível em: https://www.researchgate.net/publication/228754044_A_Memory-Based_Approach_to_Anti-Spam_Filtering. Acesso em: 04 ago. 2025.

SILVA, C. B.; GOMES, F. F. B.; SANTOS, J. V. C.; SANTOS, C. S. e. UMA ANÁLISE COMPARATIVA DAS TÉCNICAS DE MACHINE LEARNING: Regressão Logística, Árvores de Decisão, Random Forest e SVM. **Apoena**, *[S. l.]*, v. 7, p. 501–511, 2023. Disponível em: https://publicacoes.unijorge.com.br/apoena/article/view/183. Acesso em: 22 jun. 2025.

VALDATI, A. B. **Inteligência Artificial – IA**. 1. ed. Curitiba: Contentus, 2020.