XII Seminário de Extensão e Inovação
XXVII Seminário de Iniciação Científica e Tecnológica da UTFPR
Pesquisa, Extensão e Inovação: Caminhos para a Sustentabilidade
7 a 11 de novembro de 2022 – Santa Helena, PR

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS SANTA HELENA

Aprendizado de Máquina Aplicado à Detecção de Spam em E-mails

Machine Learning Applied to Email Spam Detection

Luana Cristina Guerreiro Campos[1],
Gustavo Henrique Paetzold[2]

RESUMO

Constantemente usuários da Internet sofrem tentativas de golpes através de e-mails, assim surge a necessidade de criação de filtros para bloquear spam. Aprendizado de Máquina busca automatizar o processo de reconhecimento de padrões com modelos matemáticos, desta forma, este trabalho investigou a assertividade dos principais algoritmos de aprendizado supervisionado, como classificadores Naive Bayes, Árvores de Decisão, Florestas Aleatórias e Máquina de Vetores de Suporte. A base de dados utilizada foi retirada da plataforma consolidada Apache SpamAssassin, e os textos foram processados com o algoritmo de Porter Stemmer no qual realiza a extração dos radicais de palavras e também transformados em forma numérica através da técnica de Frequência do Termo - Frequência Inversa do Termo. Por fim, o experimento que obteve melhor resultado, atingiu 98,20% e 99,23% de acurácia e precisão, respectivamente. A programação contou com a linguagem de programação Python e o auxílio de bibliotecas como Pandas, Scikit-learn e Natural Language Toolkit.

PALAVRAS-CHAVE: Aprendizado de Máquina; Detecção de Spam; Processamento de Linguagem Natural.

ABSTRACT

Internet users are constantly suffering scam attempts through e-mails, so there is a need to create filters to block spam. Machine Learning search to automate the pattern recognition process with mathematical models, in this way, this work investigated the assertiveness of the main supervised learning algorithms, such as Naive Bayes, Decision Trees, Random Forests and Support Vector Machine classifiers. The database used was taken from the consolidated Apache SpamAssassin platform, and the texts were processed with Porter Stemmer's algorithm that extracts word's stems and also transformed into numerical form using the Term Frequency - Inverse Document Frequency technique. Finally, the experiment that obtained the best result, reached 98,20% and 99,23% of accuracy and precision, respectively. The programming relied on the Python programming language and the assistance of libraries such as Pandas, Scikit-learn and Natural Language Toolkit.

KEYWORDS: Machine Learning; Spam Detection; Natural Language Processing

[1] Bolsista da UTFPR. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: luanacampos@alunos.utfpr.edu.br. ID Lattes: 4780074821595493.

[2] Docente de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: ghpaetzold@utfpr.edu.br. ID Lattes: 3576463426605379.

INTRODUÇÃO

Com a popularização da Internet, e-mails chegaram para suprir a necessidade de comunicação, contudo, também iniciaram-se as tentativas de golpes por esse meio. E-mails maliciosos e legítimos são denominados como spam e ham, respectivamente. Estima-se que pelo menos metade dos e-mails recebidos por um usuário comum sejam spam (DADA et al., 2019). Dessa forma, é necessário a criação de ferramentas que sejam capazes de filtrar spam.

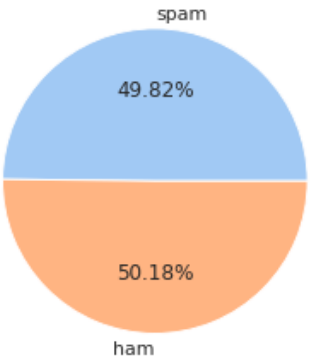
Aprendizado de máquina é uma aplicação de Inteligência Artificial capaz de aprender automaticamente através de conjuntos de dados, o aprendizado se dá através de diversas técnicas, seja criando modelos matemáticos ou estruturas tabulares. Neste trabalho analisou-se os principais algoritmos de aprendizado supervisionado, como por exemplo, Naive Bayes, Árvores de Decisão, Florestas Aleatórias e Máquinas de Vetores de Suporte.

Os e-mails foram processados com técnicas de Stemização e Vetorização TF-IDF e em seguida realizou-se o estudo de assertividade para cada método de aprendizado, utilizando bibliotecas como *Scikit-learn*, *Pandas*, *Natural Language Toolkit* e *Matplotlib*.

MATERIAIS E MÉTODOS

Apache SpamAssassin é uma plataforma anti-spam de código aberto que oferece um filtro para classificar e-mails e bloquear spam. Neste trabalho utilizou-se o conjunto de dados disponibilizado pela plataforma, sendo composto por 2790 e-mails. A Figura 1 mostra a distribuição de e-mails ham e spam. A divisão do conjunto de dados compõe-se em 80-20% para treinamento e testes, respectivamente.

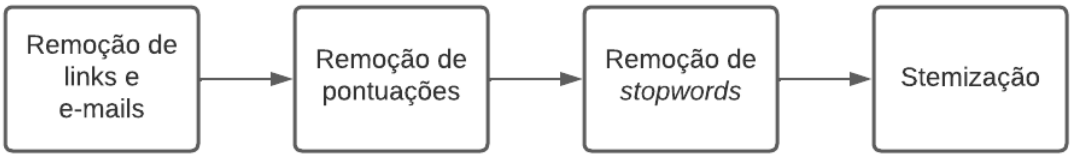
Figura 1 - Distribuição de e-mails



Fonte: Autoria própria (2022).

O corpo do e-mail foi dividido em palavras, também denominadas como *tokens*. Os *tokens* então foram processados como mostra a Figura 2.

Figura 2 - Fluxograma de processamento



Fonte: Autoria própria (2022).

A remoção de links, e-mails e pontuações foi feita por se tratar de variáveis que mesmo após sua remoção ainda seja possível a compreensão do e-mail. Como também *stopwords,* ou, palavras de parada, como por exemplo: as, e, os, de, para, com, sem e foi.

O processo de stemização consiste em reduzir uma palavra ao seu radical, por exemplo, as palavras amigos e amigas possuem a mesma raiz: amig. Dessa forma, é possível a redução da quantidade de palavras sem perder o significado semântico da frase em análise. A tarefa de stemização foi realizada através do algoritmo de *Porter Stemmer*, desenvolvido para remoção de sufixos (PORTER, 2006).

A vetorização de um texto se dá em transformá-lo em forma numérica, utilizou-se a técnica de Frequência do Termo - Frequência Inversa do Termo, ou em inglês *Term Frequency — Inverse Document Frequency* (TF-IDF) que baseia-se na frequência das palavras em um documento como também a sua relevância.

Algoritmos de aprendizado de máquina supervisionados aprendem a partir de um conjunto de dados já rotulados com a saída esperada, nesse contexto, tem-se quais e-mails são spam ou não. São os algoritmos mais utilizados:

NAIVE BAYES

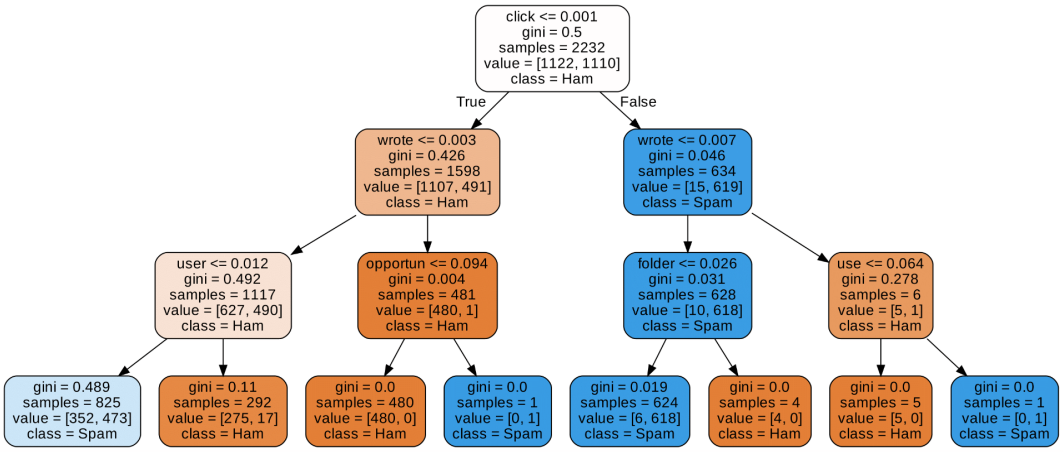
Um classificador Naive Bayes é um classificador probabilístico baseado na aplicação do Teorema de Bayes, no qual fornece a probabilidade de um evento com base nas novas informações que estão relacionadas ao evento (RAZA, 2021). Esse classificador é popular para filtros de e-mails já que funciona correlacionando as palavras que estão presentes em spams e dessa forma calcula a probabilidade do e-mail ser spam ou não.

ÁRVORES DE DECISÃO

Os classificadores de Árvores de Decisão utilizam uma série de regras para tomar decisões e determinar a qual classe a entrada pertence. No processo de treinamento as Árvores são construídas a partir de nós, e a cada nó uma das *features* é analisada através de métricas como *Gini Impurity* ou entropia a fim de dividir o conjunto de dados sucessivamente, até que todas as entradas estejam isoladas em sua classe de saída (KADHIM, 2019).

Dessa forma, torna-se fácil a tarefa de classificar novas amostras já que basta realizarmos uma busca na árvore gerada pelo treinamento, como mostra o exemplo da Figura 3.

Figura 3 - Árvore de Decisão



Fonte: Autoria própria (2022).

A medida de *Gini Impurity* indica a probabilidade de classificações erradas do nó observado, é natural assumir que árvores com maior profundidade irão amenizar essa métrica, entretanto, árvores com alta profundidade tendem a *overfitting*, ou seja, a árvore estará perfeitamente encaixada para os dados de treinamento, todavia, não há garantia de corretude para dados nunca vistos. Uma única árvore geralmente não gerará boas predições, contudo, quando múltiplas árvores são combinadas se obtém bons resultados.

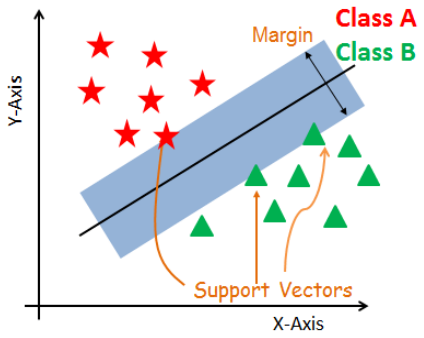
FLORESTA ALEATÓRIA

Múltiplas Árvores de Decisão quando combinadas denominam-se como Florestas Aleatórias, pois utiliza a aleatoriedade das *features* para criar uma Floresta de Árvores não correlacionadas. Enquanto uma única árvore considera todas as *features* para divisão das amostras, as Árvores da Floresta consideram apenas subconjuntos de *features*. Dessa forma, a classificação se dá através da combinação dos resultados de cada Árvore da Floresta.

MÁQUINA DE VETORES DE SUPORTE

A Máquina de Vetores de Suporte busca encontrar o hiperplano em um espaço N-dimensional que separa as amostras do conjunto de dados, para separá-las há muitas possibilidades de hiperplanos que podem ser escolhidos. O hiperplano ótimo é aquele que possui a maior margem entre os pontos das duas classes, assim, novas amostras serão classificadas com maior confiança (KADHIM, 2019). A Figura 4 exemplifica a ideia.

Figura 4 - Exemplo de hiperplano

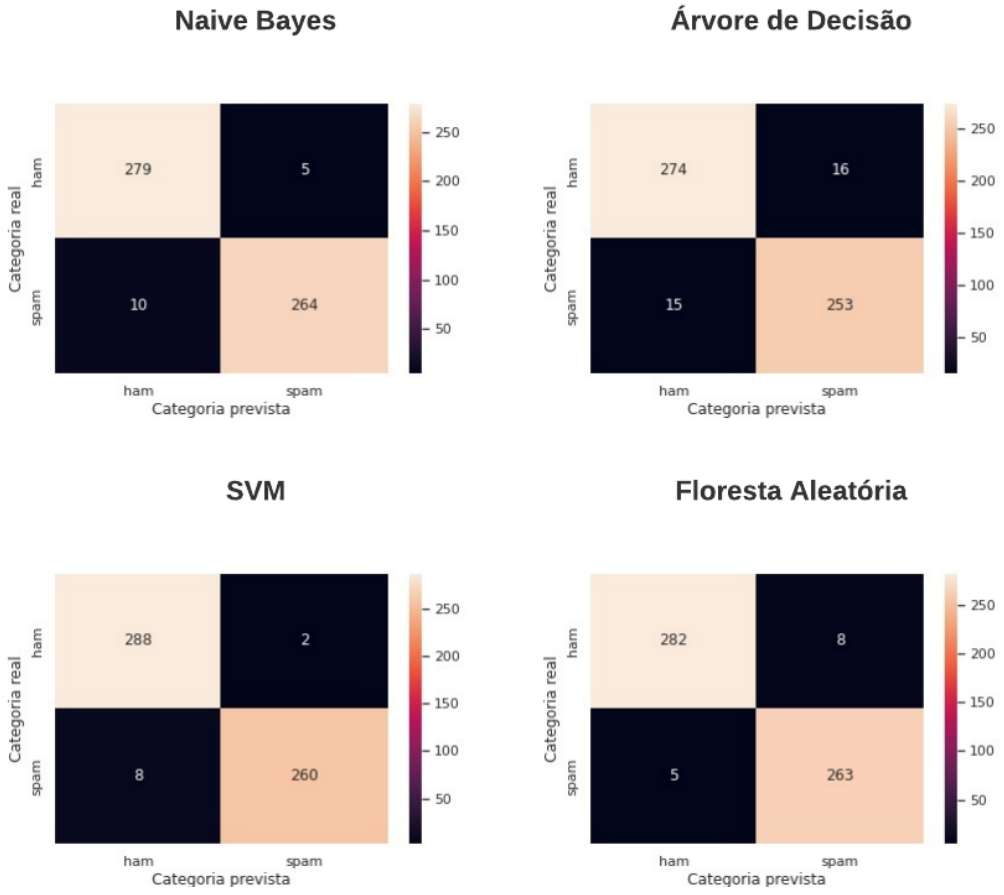


Fonte: Datacamp (2019).

RESULTADOS

A matriz de confusão nos aponta de forma direta o resultado de cada um dos algoritmos, pois mostra a quantidade de acertos de cada classe como também falsos negativos e falsos positivos. Cada uma das matrizes estão descritas na Figura 5.

Figura 5 - Matrizes de confusão



Fonte: Autoria própria (2022).

CONCLUSÃO

Usuários sofrem com tentativas de golpes através de e-mails, assim surge a necessidade de criar e aprimorar sistemas inteligentes capazes de filtrar e-mails maliciosos. Neste trabalho investigou-se os principais métodos de Aprendizado de Máquina para classificação de e-mails, como Naive Bayes, Árvores de Decisão, Floresta Aleatória e Máquina de Vetores de Suporte. Também foi apresentado as técnicas utilizadas para processamento de textos, como o algoritmo de Porter Stemmer. Por fim, foi possível analisar a assertividade de cada um dos métodos de Aprendizado de Máquina, sendo o melhor deles, a Máquina de Vetores de Suporte com 98,20% e 99,23% de acurácia e precisão, respectivamente. O presente trabalho foi desenvolvido com a linguagem de programação Python e bibliotecas como *Pandas, Scikit-learn e Natural Language Toolkit*.

Agradecimentos

Agradeço a UTFPR por fomentar essa pesquisa, como também ao professor Gustavo por me orientar.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

RAZA, M.; JAYASINGHE, N. D.; MUSLAM, M. M. A. **A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms**. 2021 International Conference on Information Networking (ICOIN), 13 jan. 2021.

DADA, E. G. et al. **Machine learning for email spam filtering: review, approaches and open research problems**. Heliyon, v. 5, n. 6, p. e01802, jun. 2019.

PORTER, M. F. **An algorithm for suffix stripping. Program,** v. 40, n. 3, p. 211–218, jul. 2006.

KADHIM, A. I. **Survey on supervised machine learning techniques for automatic text classification**. Artificial Intelligence Review, 19 jan. 2019.