Classificação de Mensagens como Spam ou Não-Spam

Palavras chave—modelo, regressão logística, classificação, spam, email (*key words****)***

I. DATASET UTILIZADO PARA CLASSIFICAÇÃO

Neste trabalho, foram usados três datasets de emails rotulados, provenientes de diferentes fontes do Kaggle, combinados em um único dataset para realizar a classificação de emails como spam ou não-spam. Os datasets utilizados foram: Spam Emails Dataset [1], Spam Mails Dataset [2], e Email Spam Classification Dataset [3].

O uso de tais datasets é justificado pela sua aplicação prática no contexto de filtragem de emails, uma área crucial para empresas que oferecem serviços de email e proteção contra mensagens maliciosas ou indesejadas. O artigo de El Idrissi et al. [4] explora a aplicação desses tipos de datasets em sistemas de classificação de emails para aprimorar o desempenho dos filtros de spam, destacando a importância de melhorar a precisão e a eficiência desses modelos para reduzir o número de falsos positivos e negativos.

II. PIPELINE DE CLASSIFICAÇÃO

Foi implementada uma pipeline de pré-processamento para classificar emails como spam. O texto foi limpo, mantendo apenas letras, números e símbolos recorrentes em spam, como sinais de moeda e "!" ou "?".

Após isso, os emails foram vetorizados com foco em termos comuns a mensagens de spam, como palavras ligadas a promoções. Contudo, a frequência de palavras pode gerar falsos positivos, já que termos ambíguos, como "desconto", também aparecem em emails legítimos. Além disso, a ausência de análise semântica dificulta a interpretação de contextos, como frases negativas.

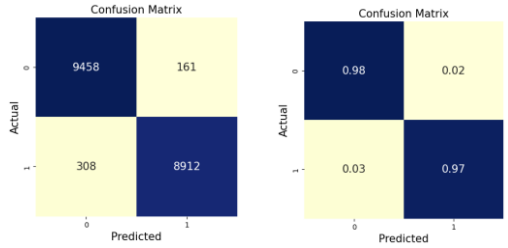
III. EXECUTANDO E AVALIANDO O MODELO

O classificador escolhido foi um modelo de regressão logística. A divisão do conjunto de dados foi realizada utilizando uma proporção de 80% para treino e 20% para teste, com o intuito de garantir uma amostra representativa para a avaliação do modelo. Essa divisão foi feita de forma aleatória em múltiplas execuções, assegurando que o modelo fosse testado em diferentes distribuições dos dados.

O modelo, após treinado, atingiu uma acurácia balanceada de 97,49%, uma métrica que ajusta o cálculo da acurácia para lidar com um possível desequilíbrio entre as classes. Esse resultado sugere que o classificador apresentou um desempenho robusto, conseguindo distinguir com precisão os emails de spam dos emails legítimos.

Para investigar a importância das palavras no processo de classificação, foram extraídos os coeficientes das palavras do modelo. Os coeficientes indicam a influência de cada termo na previsão, onde valores positivos indicam uma tendência de ser spam e valores negativos indicam uma tendência de ser não spam. As palavras com maior peso negativo, associadas à categoria não spam, incluem “enron”, “ltgt” e “vince”, que frequentemente aparecem em comunicações legítimas. Por outro lado, as palavras com maior peso positivo, associadas à categoria spam, incluem “adf”, “attach”, “medications” e “mortgage”, que comumente surgem em emails de propaganda ou fraudes financeiras.

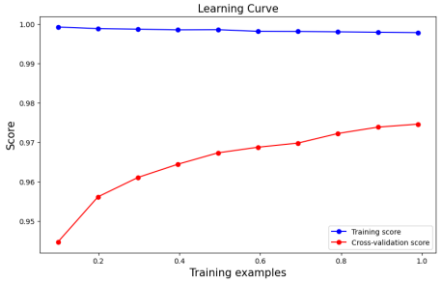
Por fim, foi verificada a matriz de confusão para avaliar o desempenho do classificador em termos de erros de classificação. O modelo foi capaz de identificar corretamente a maioria das instâncias, com poucos falsos positivos e falsos negativos. Observe as figuras abaixo:



IV. AVALIANDO O CONJUNTO DE DADOS

Para avaliar o impacto do tamanho do conjunto de dados no desempenho do modelo, foi realizada uma curva de aprendizado variando a proporção dos dados de treino de 10% a 99%. A acurácia balanceada foi medida tanto nos dados de treino quanto nos de teste, conforme o tamanho do conjunto de treinamento foi aumentando.

Os resultados indicam que a acurácia nos dados de treino permaneceu alta e estável, próxima de 1.0, o que sugere que o modelo está memorizando os dados, levando a um possível overfitting. Isso é evidenciado pelo fato de que o desempenho nos dados de treino é significativamente maior do que nos dados de teste. Observe a curva de aprendizado abaixo:



Por outro lado, a acurácia nos dados de teste aumentou consistentemente à medida que mais dados de treino foram utilizados, passando de 94,47% para 97,45%. Isso mostra que o modelo está generalizando melhor com o aumento dos dados de treino, o que é um indicativo de que o uso de mais dados pode melhorar a performance.

V. MODELOS TÓPICOS PARA REFINAR RESULTADOS

A técnica de Latent Dirichlet Allocation (LDA) foi utilizada para identificar os principais tópicos no conjunto de dados. O modelo foi configurado para encontrar 5 tópicos, e as palavras mais representativas em cada um deles foram listadas. No Tópico 0, palavras como "source", "samba" e "http" se destacaram, enquanto o Tópico 1 apresentou termos como "is", "it" e "you". Os outros tópicos seguiram um padrão semelhante, refletindo diferentes áreas temáticas no conjunto de dados.

REFERENCES

[1] A. Wagih, Spam Emails Dataset, Kaggle. Disponível em: https://www.kaggle.com/datasets/abdallahwagih/spam-emails

[2] V. Shankaranarayana, Spam Mails Dataset, Kaggle. Disponível em: https://www.kaggle.com/datasets/venky73/spam-mails-dataset

[3] P. Singhvi, Email Spam Classification Dataset, Kaggle. Disponível em: https://www.kaggle.com/datasets/purusinghvi/email-spam-classification-dataset

[4] S. El Idrissi et al., "Spam Email Classification using Machine Learning," European Journal of Electrical and Computer Engineering, vol. 5, no. 4, pp. 12-19, 2021. Disponível em: https://ejece.org/index.php/ejece/article/view/409