Classifying sentiment on the TeePublic$_{r}eviewdataset$

Guilherme dos Santos Martins
Insper
São Paulo, Brasil
guilhermesm9@al.insper.edu.br

***Index Terms*—Sentiment Analysis, TeePublic_review, Natural Language Processing (NLP), Logistic Regression, Text Classification, TF-IDF, Machine Learning, Data Preprocessing, Confusion Matrix, Topic Analysis**

I. Introdução

Este trabalho tem como objetivo a classificação de sentimentos em um conjunto de dados composto por avaliações de clientes da TeePublic, uma plataforma online voltada para itens de moda. O dataset contém textos fornecidos pelos usuários, que expressam suas opiniões e sentimentos sobre diferentes produtos. A partir de técnicas de Processamento de Linguagem Natural (PLN) e algoritmos de classificação, este estudo visa categorizar as avaliações como positivas ou negativas, fornecendo dados úteis para a melhoria de produtos e para a formulação de estratégias de marketing.

II. Dataset

O conjunto de dados utilizado nesta pesquisa é denominado "TeePublic_review" [1] e contém feedbacks detalhados dos clientes, servindo como a principal fonte para a classificação de sentimentos realizada neste estudo.

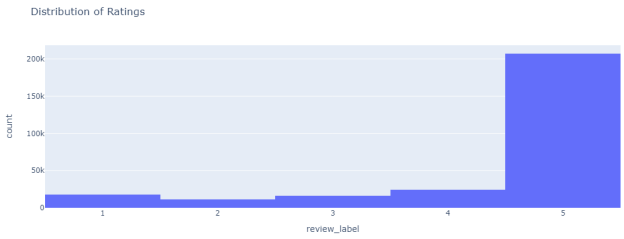

Fig. 1. Distribuição de Avaliações

A análise de distribuição dos *ratings* apresentados no gráfico acima revela uma predominância marcante de avaliações classificadas com a nota máxima, representada pelo valor **5**. Essa concentração indica que a maioria dos usuários da plataforma tendem a fornecer avaliações altamente positivas. Avaliações intermediárias, com notas entre **1** e **4**, são praticamente inexistentes no conjunto de dados, o que pode sugerir alguns aspectos a serem investigados:

1) **Tendência de Feedback Positivo**: A forte prevalência de notas **5** pode refletir uma inclinação dos clientes em expressar satisfação quando deixam feedbacks, ignorando experiências neutras ou ligeiramente negativas. Esse comportamento pode ser comum em plataformas de comércio eletrônico, onde a tendência dos usuários é avaliar positivamente apenas quando muito satisfeitos com o produto.
2) **Sub-representação de Sentimentos Negativos**: A ausência de notas mais baixas, como **1** e **2**, pode indicar uma sub-representação de clientes insatisfeitos, o que sugere a necessidade de investigar se os usuários com experiências negativas estão deixando de fornecer feedback ou se o processo de coleta de dados não captura bem essas avaliações.
3) **Possível Viés na Coleta de Dados**: O fato de o dataset apresentar uma distribuição tão desequilibrada pode indicar um viés na forma como as avaliações são coletadas ou publicadas. Talvez haja incentivos para que os clientes forneçam apenas avaliações positivas, ou pode haver uma exclusão ou subnotificação de avaliações com notas menores.

1) Impacto na Classificação de Sentimentos: Essa distribuição desigual pode afetar a acurácia dos algoritmos de classificação de sentimentos empregados no estudo. Como há uma clara concentração de avaliações positivas, os modelos podem acabar enviesados, identificando predominantemente sentimentos positivos e subestimando ou ignorando sentimentos negativos. Esse desequilíbrio de classes é um desafio clássico em tarefas de aprendizado de máquina, e estratégias como *oversampling* de classes menores ou o uso de técnicas de balanceamento, como *SMOTE* (Synthetic Minority Over-sampling Technique), podem ser necessárias para garantir uma classificação mais precisa.

Além disso, esta distribuição enfatiza a importância de ajustar os parâmetros de avaliação do modelo (como a métrica F1 ou AUC) para garantir que tanto os sentimentos positivos quanto os negativos sejam corretamente classificados.

III. Pipeline de Classificação

O pipeline de classificação segue uma abordagem estruturada, composta pelas seguintes etapas:

1) **Pré-processamento de Texto**: O texto das colunas *review* e *title* é convertido para letras minúsculas, removendo números, caracteres especiais e *stopwords*.

Aplicam-se **stemming** e **lemmatização** para normalizar as palavras.

2) **Vetorização TF-IDF**: As colunas de texto são transformadas em vetores numéricos usando a técnica **TF-IDF** (Term Frequency-Inverse Document Frequency), que pondera a frequência e a relevância das palavras.
3) **Classificação**: Um modelo de **Regressão Logística** com 1000 iterações é utilizado para classificar os sentimentos como positivos ou negativos.
4) **Pipeline**: O processo completo de pré-processamento, vetorização e classificação é encapsulado em um **pipeline**, garantindo execução eficiente e integrada.

Esse pipeline foi ajustado usando os dados de treino `X_train` e `y_train`, possibilitando a classificação eficiente de novos dados.

IV. Avaliação do Modelo

A matriz de confusão e os relatórios de classificação demonstram o desempenho do modelo em cinco classes.

A. Matriz de Confusão

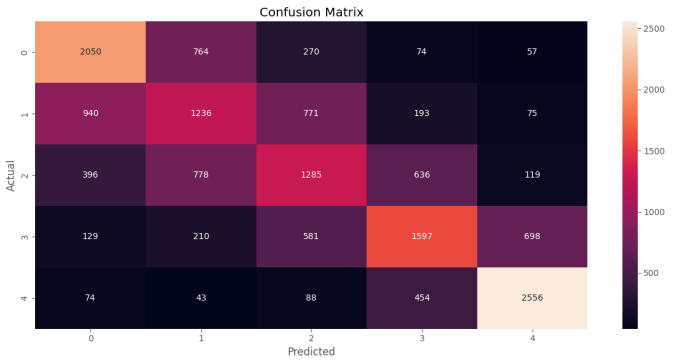

Fig. 2. Matrix de confusão entre dados reais e predições

- A classe **4** obteve o melhor desempenho, com **80%** de acertos.
- As classes **0**, **1**, **2** e **3** apresentaram alta taxa de confusão, especialmente entre as classes intermediárias.
- A maior confusão ocorreu entre as classes **1** e **2**, dificultando a distinção entre avaliações intermediárias.

B. Relatório de Classificação

a) Conjunto de Treino::

- **Acurácia** de **74%**.
- Melhor desempenho na classe **4** com **82%** de precisão e **89%** de recall.
- As demais classes tiveram precisão e recall variando de **65%** a **80%**.

b) Conjunto de Teste::

- **Acurácia** de **54%**.
- Melhor desempenho na classe **4**, com **73%** de precisão e **80%** de recall.
- Classes **1** e **2** tiveram baixo desempenho, com **f1-scores** de **0.40** e **0.41**, respectivamente.

C. Conclusão

O modelo classifica bem avaliações positivas, mas tem dificuldades em distinguir avaliações intermediárias e negativas. Melhorias no balanceamento de classes e ajuste de hiperparâmetros podem trazer uma melhora nos resultados.

V. Discussão Sobre Tamanho do Dataset

À medida que o tamanho do conjunto de treino aumentou, houve uma melhora na acurácia do modelo. No entanto, essa melhora foi limitada, com ganhos de desempenho que não se mostraram expressivos. O aumento da acurácia foi perceptível até aproximadamente 70% do conjunto de treino, momento em que os benefícios adicionais de incluir mais dados começaram a se estabilizar.

Essa estabilização sugere que, embora o modelo se beneficie de mais dados, sua capacidade de generalização pode estar limitada por outros fatores, como a necessidade de ajustes nos hiperparâmetros ou melhorias na complexidade do modelo. Isso indica que, após certo ponto, simplesmente aumentar o tamanho do dataset não é suficiente para obter melhorias significativas na performance do modelo.

Portanto, enquanto o aumento do conjunto de treino foi útil para melhorar a acurácia, essa estratégia sozinha pode não ser suficiente para alcançar um desempenho ideal, sendo necessário considerar ajustes adicionais no pipeline de classificação.

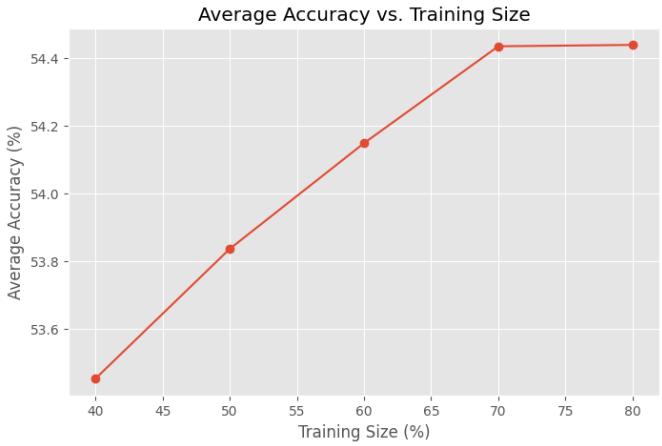

Fig. 3. Acurácia x Tamanho

VI. Análise por Tópicos

A análise por tópicos visa avaliar o desempenho do modelo em diferentes categorias ou temas, permitindo identificar variações na acurácia conforme o conteúdo do texto. Isso pode revelar onde o modelo se sai melhor ou onde encontra mais dificuldades.

1) Resultados da Análise: Os resultados mostram uma variação significativa na acurácia do modelo entre os diferentes tópicos:

- O tópico **0** obteve a maior acurácia, atingindo **50%**, indicando que o modelo se sai relativamente bem ao classificar exemplos desse tema.

- O tópico **1** apresentou o pior desempenho, com acurácia abaixo de **40%**, sugerindo que o modelo encontra dificuldades em classificar corretamente avaliações desse tópico.
- Os tópicos **2** e **4** tiveram um desempenho intermediário, enquanto o tópico **3** se destacou com uma acurácia ligeiramente superior a **50%**.

VII. Conclusão

O modelo de classificação mostrou um desempenho moderado, com variações significativas em função do tópico. Embora o aumento no tamanho do dataset tenha proporcionado melhorias na acurácia, elas não foram substanciais. A análise por tópicos revelou que o modelo se sai bem em alguns temas, mas encontra dificuldades em outros. Para otimizar o desempenho geral, recomenda-se explorar ajustes no modelo e o balanceamento de classes, além de testar abordagens específicas para melhorar a classificação em tópicos com acurácia mais baixa.

References

[1] ShopperSentiments. [Online]. Disponível em: *https://www.kaggle.com/datasets/nelgiriyewithana/shoppersentiments*

[2] ensemble_classification_model. [Online]. Disponível em: *https://www.kaggle.com/code/estiven0507/ensemble-classification-model*