Classificando Discurso de ódio e linguagem ofensiva no dataset Hate Speech and Offensive Language Dataset

Gabriel Valentim
Insper
São Paulo, Brasil
joaogvr@al.insper.edu.br

I. Dataset

Neste estudo, utilizamos o *Hate Speech and Offensive Language Dataset* disponível no *Kaggle* e originalmente apresentado em [1]. O dataset é composto por tweets classificados em três categorias: discurso de ódio, linguagem ofensiva e linguagem neutra. O objetivo é desenvolver um modelo capaz de identificar automaticamente conteúdos ofensivos em redes sociais, contribuindo para a moderação de conteúdo e proteção dos usuários.

II. Pipeline de Classificação

Implementamos uma *pipeline* de processamento que inclui:

- **Pré-processamento**: Remoção de *stopwords*, pontuações, números e palavras curtas; lematização para reduzir as palavras à sua forma básica.
- **Vetorização**: Utilização do *TF-IDF Vectorizer* para transformar o texto em uma representação numérica.
- **Classificação**: Modelo de Regressão Logística [2], onde os coeficientes indicam a importância de cada palavra na classificação.

Para simplificar o problema, convertemos a tarefa em uma classificação binária, unindo as classes de discurso de ódio e linguagem ofensiva em uma única classe ofensiva. Também aplicamos *undersampling* para equilibrar as classes.

III. Avaliação

Utilizamos validação cruzada com 100 *folds* para obter estimativas estáveis das métricas de desempenho. As métricas avaliadas foram acurácia, precisão, revocação e F1-Score. A média e o desvio padrão das métricas foram calculados, e os resultados mostram um desempenho consistente do modelo, que foi registrado na Figura 1.

A. Análise das Palavras Mais Importantes

Diante do modelo utilizado, foi possível ordenar em uma lista as palavras que mais impactam a inferência do modelo, tanto no contexto positivo quanto negativo.

- Contribuições Positivas: As palavras contribuindo positivamente fizeram sentido, dado seu caráter neutro. Isso mostra indícios de que o modelo está de fato utilizando as palavras corretamente para a inferência.
- Contribuições Negativas: As palavras contribuindo negativamente também fizeram sentido, dado que a maioria delas são claramente de linguagem ofensiva. O que também mostra que palavras ofensivas estão corretamente associadas à classe negativa.

Diante disso, exploramos também casos nos quais o classificador pudesse ter dúvida ordenando as instâncias pela entropia da classificação. Isso apontou para casos em que o modelo pode falhar como **uso de palavras ofensivas em contextos neutros** ou **incerteza na polaridade emocional**.

IV. Tamanho do Dataset

Analisamos o impacto do tamanho do *dataset* nas curvas de aprendizado geradas pelo modelo, variando a proporção de dados de treinamento como pode ser visto na Figura 2. Conforme o tamanho do *dataset* aumenta, o erro de treinamento aumenta enquanto o erro de teste tende a diminuir, convergindo para um valor assintótico. Inicialmente, o erro de teste é significativamente maior que o de treinamento, porém conforme mais dados são utilizados para o treinamento, ambos os erros se aproximam. Por fim, com base no estudo registrado em [4], podemos construir o gráfico da Figura 3, que mostra que ainda é viável aumentar o tamanho do *dataset* e ainda ter ganhos significativos. No contexto de negócios, é viável ampliar o *dataset* utilizando APIs de redes sociais para coletar mais dados.

V. Análise de Tópicos

Aplicamos a Modelagem de Tópicos utilizando *Non-negative Matrix Factorization (NMF)* [3] para identificar tópicos latentes nos *tweets*. Atribuímos a cada documento um tópico dominante e treinamos modelos específicos para cada tópico. A avaliação revelou variações na taxa de erro entre os tópicos, indicando que o modelo tem desempenho melhor em certos contextos. O classificador de duas camadas, primeiro identificando o tópico e depois aplicando um modelo especializado, apresentou melhorias em comparação com o modelo geral, do ponto de vista das métricas citadas na Seção III.

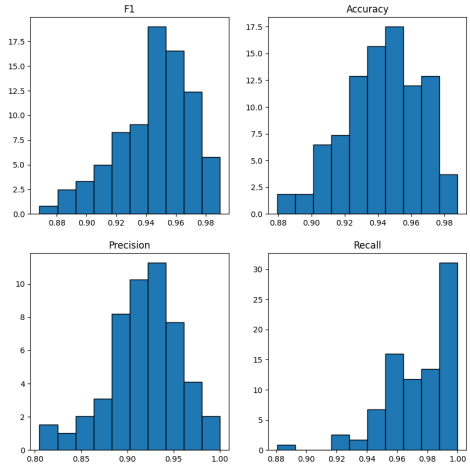

Fig. 1. A figura mostra distribuições geradas pela validação cruzada que foi aplicada no modelo. O método de *kfolds* gera as distribuições trazendo maior confiabilidade da performance do modelo ao longo das métricas que definimos para a avaliação.

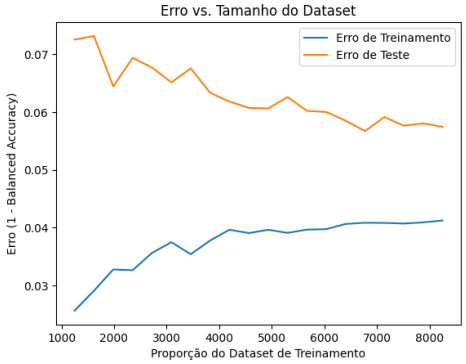

Fig. 2. A figura mostra as curvas de aprendizado geradas pelo modelo. É possível observar que inicialmente o erro de generalização começa alto e o de treino começa baixo. Porém, com o aumento do tamanho do *dataset* elas acabam convergindo para um ponto, como é registrado na Seção IV.

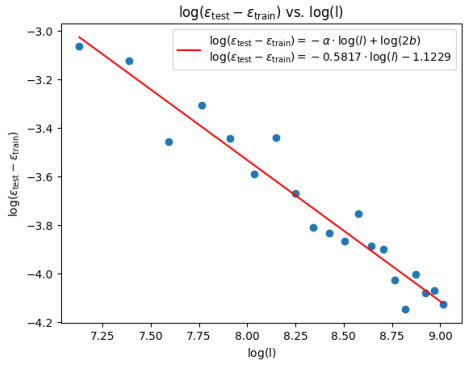

Fig. 3. A figura mostra um gráfico com o qual aplicamos as técnicas registradas em [4], que em suma mostra a viabilidade de continuar ou não aumentando o *dataset*. No nosso caso, ainda faz sentido tanto do ponto de vista técnico quanto do ponto de vista de negócios.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, 'Automated hate speech detection and the problem of offensive language," in *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2017, pp. 1–11.

[2] J. S. Cramer, 'The origins of logistic regression," Tinbergen Institute Working Paper, no. 2002-119/4, 16 pages, Dec. 2002.

[3] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," Advances in Neural Information Processing Systems, vol. 13, pp. 556-562, 2001.

[4] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, 'Learning curves: Asymptotic values and rate of convergence," in *Advances in Neural Information Processing Systems (NIPS)*, 1993, pp. 327–334.