Classificação de Tweets de Desastres no Dataset Kaggle

Rafael Evangelista Monteiro
Insper
rafaelem2@al.insper.edu.br

I. Dataset

O dataset utilizado neste projeto foi extraído da competição ”Natural Language Processing with Disaster Tweets” do Kaggle [1]. Ele contém 7.613 tweets, rotulados como relacionados a desastres (1) ou não relacionados a desastres (0). O objetivo deste dataset é classificar tweets com base em se eles descrevem ou não um evento de desastre real, o que pode ajudar os serviços de emergência a filtrar dados de mídias sociais em tempo real. O dataset contém cinco colunas: ID do tweet, palavra-chave, localização, texto e rótulo (target).

Uma análise exploratória revelou um leve desbalanceamento na variável ‘target‘, com mais tweets não relacionados a desastres. As palavras-chave mais frequentes foram termos diretamente ligados a desastres, como ”fatalities” e ”evacuate”. Além disso, os tweets relacionados a desastres tendem a ser ligeiramente mais longos.

O arquivo train.csv foi utilizado para o treinamento e a avaliação do modelo porque contém a coluna alvo (target). Essa coluna fornece a informação para que os modelos aprendam com exemplos rotulados e sejam avaliados com base em sua capacidade de classificar corretamente novos dados.

II. Pipeline de Classificação

Os dados foram pré-processados para serem limpos e preparados para a classificação. As etapas incluíram:

- **Tokenização**: Cada tweet foi dividido em palavras individuais (tokens).
- **Remoção de Stopwords**: Palavras comuns do inglês (como ”the”, ”is”, ”in”) foram removidas para reduzir o ruído.
- **Stemming**: As palavras foram reduzidas à sua forma base usando o Porter Stemmer (por exemplo, ”running” se torna ”run”).

Após o pré-processamento, as características foram extraídas usando o método de Term Frequency-Inverse Document Frequency (TF-IDF) para transformar o texto em representações numéricas adequadas para modelos de aprendizado de máquina.

Três classificadores foram utilizados: Regressão Logística, Máquina de Vetores de Suporte (SVM) e Naive Bayes.

III. Avaliação

A Tabela I resume o desempenho da classificação dos três modelos no conjunto de teste:

O Naive Bayes apresentou o melhor desempenho geral.

TABLE I
Resultados da Classificação

Modelo	Precisão	Recall	F1-Score
Regressão Logística	0.80	0.68	0.73
SVM	0.79	0.70	0.74
Naive Bayes	0.82	0.70	0.75

IV. Análise do Tamanho do Dataset

Para avaliar o impacto do tamanho do dataset, os dados foram reduzidos em diferentes níveis (10%, 25%, 50%, 75%, 99%). Como mostrado na Figura 1, a acurácia balanceada melhorou de forma constante conforme mais dados foram adicionados, com retornos decrescentes além de 75% do tamanho do dataset.

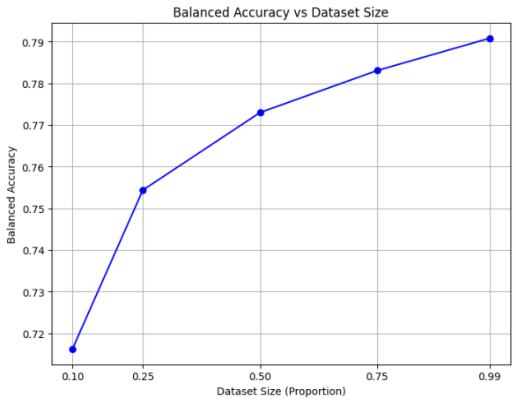

Fig. 1. Acurácia balanceada em diferentes tamanhos de dataset.

V. Conclusão

Neste projeto, utilizamos três classificadores de aprendizado de máquina para prever tweets relacionados a desastres, sendo o Naive Bayes o modelo de melhor desempenho. O train.csv foi essencial para o aprendizado supervisionado e as previsões no test.csv confirmaram a eficácia do modelo em classificar novos tweets. O projeto foi bem-sucedido em demonstrar a precisão da classificação de tweets usando técnicas de aprendizado de máquina.

References

[1] Kaggle, ”Natural Language Processing with Disaster Tweets”. Disponível em: https://www.kaggle.com/competitions/nlp-getting-started/overview.

[2] S. S. Monfared and H. Rahmati, ”Natural Language Processing for Prediction of Disaster Tweets using Machine Learning Methods”. Disponível em: https://www.researchgate.net/profile/Samane-Sharifi-Monfared/publication/352550771$_Natural_Language_Processing_for_Prediction_of_Disaster_Tweets_using_Machine_Learning_Methods/links/60cee83692851ca$
$Language - Processing - for - Prediction - of - Disaster - Tweets - using - Machine - Learning - Methods.pdf$.