Universidade Federal de Sergipe
Equipe: Adicley de Oliveira Costa, Ana Beatriz da Cunha Figueiredo, Victor Santos Chagas
Projeto: Natural Language Processing With Disaster Tweets
Link: https://www.kaggle.com/competitions/nlp-getting-started
Link notebook referência:
https://www.kaggle.com/code/sarvagyasharma10/disaster-tweet-classifier

Relatório Técnico

O desafio consiste na criação de um modelo para analisar e interpretar tweets postados sobre possíveis desastres naturais iminentes. Isto é, através do conteúdo dessas publicações, busca-se compreender o contexto e o intuito que aquilo foi escrito a fim de relacioná-lo ou não com catástrofes naturais.

Para desenvolvimento da atividade, foi realizado uma análise exploratória dos dados textuais no dataset do Kaggle para verificar a necessidade de um tratamento e limpeza dos corpus. Após a limpeza, tratamentos e conversão dos dados textuais em embeddings, foram escolhidos 3 modelos de classificação (XGBoost, SVM (Kernel RBF) e Random Forest), e, ao final, foi montado um preditor (ensembler) com base nos modelos que obtiveram maior curva ROC (taxa verdeiro positivo/taxa falsos positivos) utilizando o modelo de regressão logística como preditor final (em caso de empate) com o objetivo de criar um preditor final com melhores métricas de avaliação em relação às métricas individuais de cada um.

Tais modelos foram escolhidos pois permitem uma maior flexibilidade de ajuste nos hiperparâmetros que ajudam a adequar o modelo para os dados em questão. Para isso, foram montados gráficos que verificam a acurácia (treino vs teste) dos modelos com base na lista de valores dos hiperparâmetros escolhidos para cada um.

Com os hiperparâmetros escolhidos para cada modelo e após vários ajustes para obter melhor precisão e acurácia, o xgboost e o svm tiveram destaque pois obtiveram melhores métricas de avaliação (precision, recall, f1-score etc) nos teste, com cerca de 80% e 81% de acurácia, respectivamente, e melhores curvas ROC, com cerca de 0.83 e 0.87, respectivamente.

Após isso, foi feito o enssemble com os modelos ditos anteriormente, o qual, juntos, obtiveram uma acurácia de 82%, superando em todas as demais métricas a implementação do notebook referência utilizado para realização da atividade.

Antes disso, foram realizadas várias combinações entre os modelos e teste de hiperparâmetros, com a primeira submissão no kaggle com score de 77.6%, próxima do notebook referência que foi de 77.7%. Onde, após os ajustes de combinação dos modelos e ajustes dos hiperparâmetros, o modelo final (enssembler) apresentou score final de 79.6% ocupando a posição 297 no leaderboard do kaggle.

Por fim, o projeto pode ser adaptado para diversas tarefas de processamento de linguagem natural relacionadas a tarefas de classificação, entre elas podemos citar:

- Identificar postagens/publicações em rede social e email de cunho preconceituoso ou de incentivo à violência dentro do contexto da UFS.
- Mapeamento de zonas muito perigosas na cidade baseado nas postagens e comentários de diferentes redes sociais ligadas àquela região. Isto é, através dos comentários de determinado local, identificar o quão perigoso é aquela região. Servindo, portanto, como auxílio para pessoas que não pertencem aquela região e ou a desconhecem, auxiliando, inclusive, os próprios moradores.
- Identificação de relatórios fraudulentos e identificação de emails legítimos ou spam.