



Universidade Federal da Paraíba
Centro de Informática

Processamento de Linguagem Natural - 2023.1

Projeto final

Rebeca Andrade Medeiros - Matrícula: 11500782

João Pessoa, 2023

1. Apresentação do problema

A classificação de texto de acordo com o sentimento é uma importante tarefa para diversas áreas e o ‘Twitter’ é uma das redes sociais mais conhecidas do mundo, muito do que é postado nesta rede é compartilhado em outras redes e analisado por empresas e pesquisadores.

O ‘Twitter’ é conhecido também como uma rede onde as pessoas emitem opiniões de diversos assuntos, expressando vários tipos de sentimentos diferentes. A classificação de tweets pode ser utilizada de diversas formas, como, para avaliar opiniões sobre algum produto ou artista ou identificar mensagens de ódio.

2. Objetivos

O objetivo deste projeto era realizar a classificação de mensagens de texto da rede social ‘Twitter’, utilizando uma rede LSTM (Long Short Term Memory) para classificar dois subsets do dataset ‘tweet_eval’ e analisar as respostas obtidas de treino e teste para cada, considerando a diferença na quantidade de ‘labels’ que cada subset possui.

3. Dados utilizados e pré-processamento dos dados

O dataset escolhido para este projeto foi o ‘tweet_eval’ disponível no Hugging Face. Este dataset possui 7 subsets: irony, hate, offensive, stance, emoji, emotion, e sentiment. O subset stance está dividido em alguns assuntos.

Os subsets trabalhados neste projeto foram: emoji e emotion. Os dois possuem multiclass, emotion com 4 categorias de classificação e emoji com 20. Os dados estão separados em train, validation e test. Os dados de texto estão em inglês.

Para o pré-processamento foi realizada a criação do vocabulário para os dois subsets, transformação do texto para letras minúsculas, tokenização do texto e a transformação de token para id.

4. Rede neural

Para este projeto foi escolhida a rede neural recorrente LSTM, por conta da capacidade de lembrar informações por um maior tempo e considerando também a diversidade de tamanhos que as mensagens do 'Twitter' podem ter.