Capítulo 1

“...ferramentas computacionais capazes de processar a linguagem humana têm ganhado destaque, e a demanda pela solução de problemas reais a partir de técnicas de **Processamento de Linguagem Natural (PLN)** emergem aos montes nos mais diferentes setores da sociedade. Entende-se por Processamento de Linguagem Natural, grosso modo, **o emprego de técnicas computacionais para processar, compreender, ou ainda, gerar linguagem humana** (HAGIWARA, 2021)...”

“... Um dos maiores expoentes dessa demanda é o Exame Nacional do Ensino Médio (ENEM), visto que em 2015 o custo para a correção manual de cada redação era de 15,88 R$. Naquele mesmo ano, foram registradas 6,4 milhões de redações corrigidas, com um custo total de pelo menos 101 milhões aos cofres públicos. Havendo discrepância entre as avaliações dos dois primeiros avaliadores, um terceiro avaliador é acionado e automaticamente os custos por redação aumentam ...”

“ ... As redações são avaliadas segundo 5 critérios descritos na cartilha do participante de 20223 sendo eles:

1. Demonstrar domínio da modalidade escrita formal da língua portuguesa
2. Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
3. Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
4. Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação
5. Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos. ...”

O dataset Essay-BR (MARINHO; ANCHIÊTA; MOURA, 2021)

O Essay-BR é um corpus de ensaios dissertativos escritos em língua portuguesa cujos textos seguem o padrão mais atual do ENEM. A iniciativa se deu quando um grupo de pesquisa da Universidade Federal do Piauí (UFPI), em contribuição para os avanços nas pesquisas em Avaliação Automática de Redações (do inglês, Automatic Essays Scoring - AES), decidiu criar um corpus atualizado com textos escritos em português do Brasil. O corpus contém 4,570 documentos, divididos em 86 tópico como direitos humanos, política e cultura.

Os textos coletados foram então pré-processados, onde se removeu **tags HTML** e anotações dos avaliadores e as **pontuações foram normalizadas**, já que o site utilizava **uma estratégia diferente** da empregada pelo exame oficial.

De onde os dados foram coletadas as redações?

Sites

1. Banco de Redações UOL 1 , e o
2. Projeto Redação nota 1000