Classificação de Perguntas em Português com Processamento de Linguagem Natural

João Vitor Menezes Magalhães e Rafael Monteiro
Insper
joaovmm@al.insper.edu.br e rafaelem2@al.insper.edu.br

I. Introdução

O objetivo principal deste projeto foi classificar perguntas em português em categorias temáticas predefinidas. Essa tarefa é crucial para a organização de conteúdos educacionais, permitindo que instrutores gerenciem e avaliem materiais de forma eficiente. O conjunto de dados utilizado apresentou desafios significativos, como tamanho reduzido, desequilíbrio entre classes e necessidade de extrair características relevantes de textos curtos.

II. Definição do Problema

O projeto enfrentou desafios inerentes ao conjunto de dados, incluindo:

- Tamanho reduzido do conjunto de dados.
- Desequilíbrio entre classes.
- Complexidade na extração de características significativas de textos curtos.

III. Abordagens Utilizadas

Três abordagens foram implementadas e comparadas:

A. Bag of Words (BoW)

Uma representação clássica onde cada pergunta é transformada em um vetor de contagem de palavras. Utilizou-se um classificador Random Forest.

- **Pontos Fortes:** Fácil implementação e eficiência computacional.
- **Pontos Fracos:** Ignora a ordem das palavras e o contexto.
- **Resultados:** F1-Score: 0,41; Acurácia: 0,43.

B. TF-IDF

Utilizou-se Term Frequency-Inverse Document Frequency (TF-IDF) para ponderar a importância das palavras e um classificador Random Forest.

- **Pontos Fortes:** Penaliza palavras comuns, melhorando a distinção entre categorias.
- **Pontos Fracos:** Carece de compreensão semântica.
- **Resultados:** F1-Score: 0,43; Acurácia: 0,48.

C. Multilayer Perceptron (MLP)

Uma abordagem de deep learning com embeddings treinados do zero, utilizando camadas densas com ativações ReLU e dropout para regularização.

- **Pontos Fortes:** Modela relações não lineares.
- **Pontos Fracos:** Requer conjuntos de dados maiores ou embeddings pré-treinados.
- **Resultados:** F1-Score: 0,09; Acurácia: 0,11.

IV. Resultados

Os métodos clássicos (BoW e TF-IDF) superaram o MLP. A ligeira melhoria do TF-IDF em relação ao BoW destaca a eficácia de ponderar palavras específicas. O desempenho ruim do MLP reflete as limitações do conjunto de dados pequeno e a ausência de embeddings pré-treinados.

V. Comparação com Trabalhos Relacionados

Os resultados foram comparados com o artigo *"Aprimorando a classificação de descrições de produtos em português com a utilização de técnicas da recuperação de informação"*:

- **Pré-processamento:** Tokenização e remoção de stop words contribuíram para o desempenho superior no artigo.
- **Tamanho do Conjunto de Dados:** O artigo utilizou um conjunto de dados maior e mais equilibrado.
- **Resultados do Artigo:** F1-Scores entre 0,55 e 0,65.

Os resultados do projeto (F1-Score: 0,43) destacam a importância de conjuntos de dados maiores e pré-processamento adequado.

VI. Direções Futuras

Melhorias recomendadas incluem:

- Incorporar embeddings pré-treinados, como FastText ou Word2Vec.
- Aplicar técnicas de aumento de dados (data augmentation).
- Testar modelos baseados em Transformers, como o BERTimbau.
- Explorar estratégias de pré-processamento específicas para o domínio educacional.

VII. Conclusão

O projeto demonstrou que métodos clássicos, como BoW e TF-IDF, continuam eficazes para conjuntos de dados pequenos e textos curtos. Embeddings pré-treinados e estratégias de aumento de dados são fundamentais para liberar o potencial de abordagens modernas, como deep learning. O trabalho estabelece uma base sólida para futuras investigações combinando técnicas avançadas e pré-processamento aprimorado.

References

1. Daru GH, Loch GV, Pietezak DF. Aprimorando a classificação de descrições de produtos em português com a utilização de técnicas da recuperação de informação: uma abordagem de agrupamento de descrições. Em Quest [Internet]. 2024;30:e–139205. Available from: https://doi.org/10.1590/1808-5245.30.139205