ENAP Bootcamp

Assistente Virtual Agro

Aplicação de Machine Learning na classificação de temas e culturas como base para a criação de um assistente virtual.

Wellington Rangel - EMBRAPA

Tópicos

01 **Problema**

Descrição do problema

02 **Solução**

Solução e abordagem técnica escolhida

03 **Dados/Treinamento**

Pré-processamento treinamento e métricas

04 **Considerações finais**

Limitações e trabalhos futuros

Problema

- A Embrapa dispõe de uma série de livros com perguntas e respostas em vários temas;
- Além do formato PDF, podemos ter outras alternativas para popularizar ainda mais o acesso ao conteúdo.

Solução

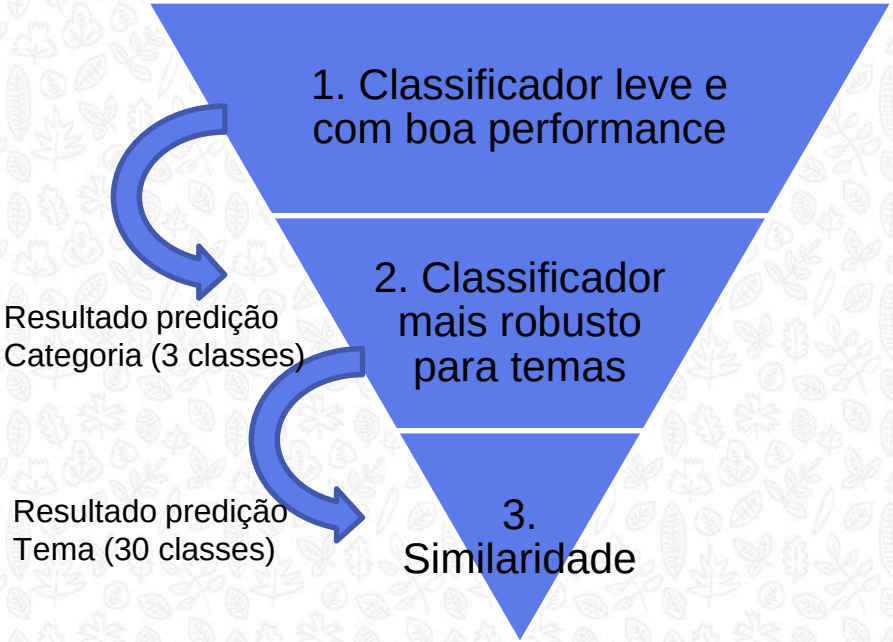
Geral

- Facilitar o acesso a esta informação;
- Entrar o conteúdo em formato mais amigável e interativo;
- Registrar o interesse da sociedade em cada tema;
- Manter canal com cada cidadão consumidor deste serviço.

Específico

- Aplicar aprendizado de máquina para prever a categoria entre as perguntas presentes nos livros;
- Combinar técnicas diferentes para entregar a melhor resposta dada uma pergunta feita em um dos temas.

Abordagem



- Diante da complexidade da tarefa e pouco dado capaz de prever a intenção das perguntas, aplicamos:

1. Classificador para prever as categorias: Frutas, Animais e Grãos
2. Classificador mais especializado em prever temas das perguntas;
3. Buscador de respostas por similaridade das perguntas em banco de dados específico e filtrado pelos modelos.

Dados

Descrição

- Os conteúdos em pdf foram transformado em um dataset (16.592 observações) com Categoria, Tema, Pergunta e Resposta;
- São três categorias: Frutas (7035), Animais (4029) e Grãos (5532);
- São 30 Temas diversos temas como: Banana, Caju, Gado, Suínos, Arroz e Soja.

Preprocessamento

- Alguns temas foram unificados como: Gado de corte, Gado no pantanal e Gado de leite para o label “Gado”;
- Aplicamos caixa baixa, caracteres especiais, stopwords, stemming, corretor ortográfico e tfidf;
- Balanceamento dos dados com RandomOverSampler;

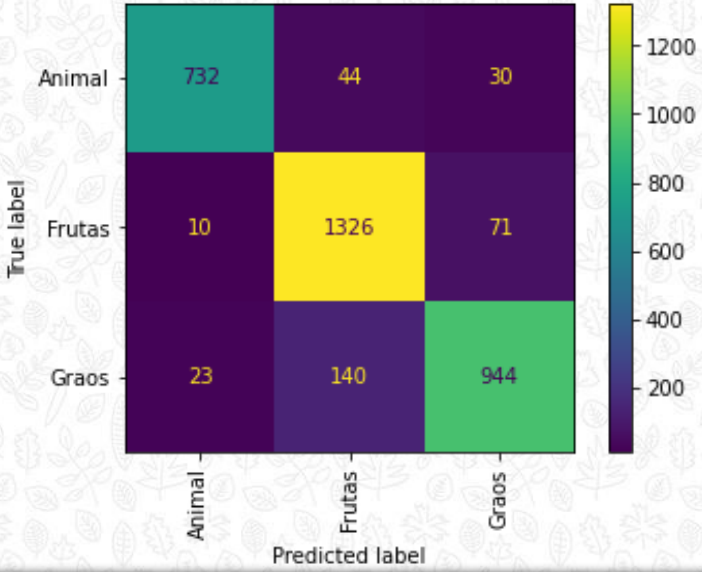
Modelo 1 – target Categoria

- Começamos com um baseline em modelo para entender os dados;
- Testamos ensemble e Gridsearch;
- As métricas foram acurácia e f1-score;
- O balanceamento de dados com RandomOverSampler não mudou a acurácia;
- Ensemble Learning teve desempenho similar ao melhor modelo.

GridSearch	Acurácia
KNeighbors	0.77
DecisionTree	0.82
MultinomialNB	0.87
RandomForest	0.83
SGDClassifier	0.90

- **word2vec** (acurácia 0.65).

Modelo 1 – SGDClassifier



- Há maior erro entre as classes Frutas e Grãos;
- Os menores erros estão relacionados a classe Animal;
- Observamos algumas perguntas semelhantes entre as classes Frutas e Grãos.

Treinamento

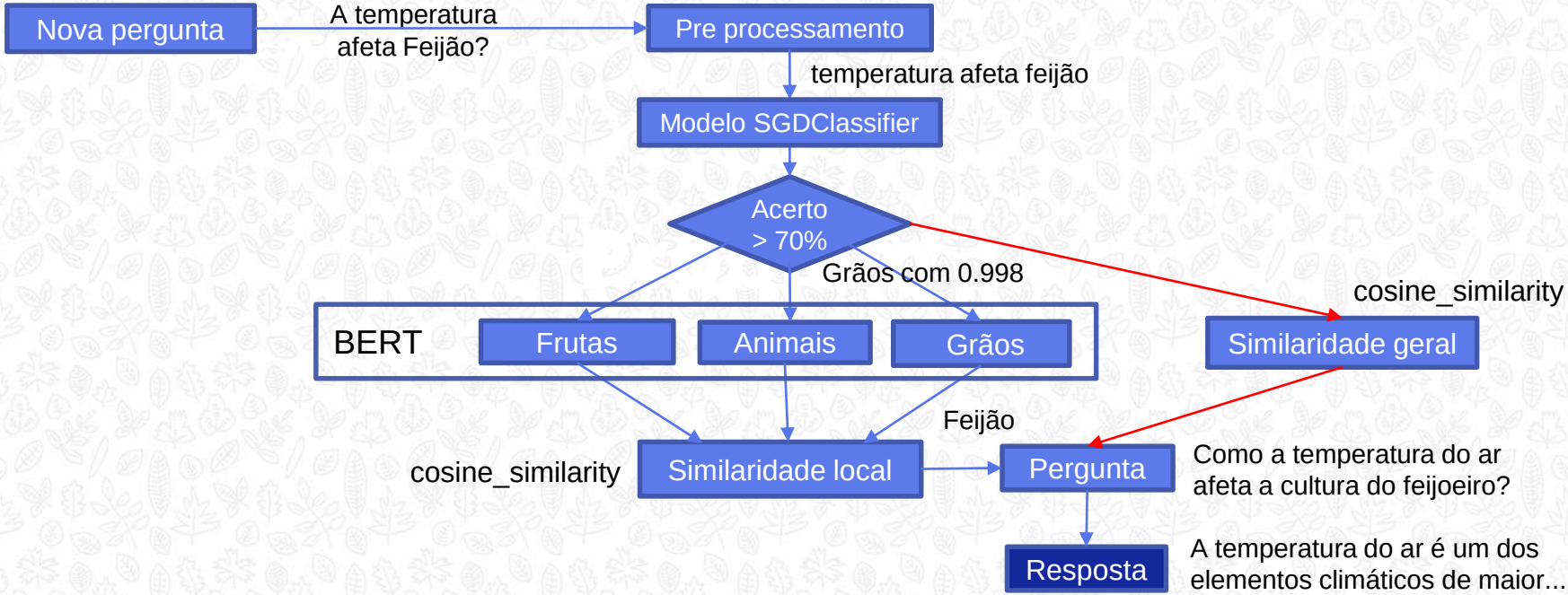
Modelo 2 – target Tema

- Modelo Baseline foi SGDClassifier
- Treinamos **3 modelos** com BERT para prever os temas das categorias Frutas, Animais e Grãos;
- Médias de 2 horas para o treinamento usando BERT com GPU.

Performance

- SGDClassifier - Acurácia (F=0.71, G=071 e A=0,77);
- BERT:
  - Temas da cat. Frutas: acurácia - 0.83; f1score - 0.70);
  - Temas da cat. Grãos: acurácia - 0.90; f1score - 0.76);
  - Temas da cat. Animais: acurácia - 0.93; f1score - 0.84);

Orquestração dos modelos



Instância do assistente

```
TESTE:
A falta de chuva prejudica a plantação do abacaxi?
['falta chuva prejudica plantação abacaxi?']
SGDClassifier
Frutas
0.9985018893602181
BERT:
ML:
Categoria: Frutas
Tema: Abacaxi
Pergunta: A falta de chuva prejudica o abacaxizeiro?
Resposta: Dependendo do estágio de desenvolvimento da cultura, a fa
```

```
TESTE:
O feijão é bom para a saúde?
['feijão bom saúde?']
SGDClassifier
Graos
0.9960814385067948
BERT:
ML:
Categoria: Graos
Tema: Feijao
Pergunta: Por que o feijão é um alimento tão bom para a saúde?
Resposta: O feijão é bom para a saúde porque ele fornece carboidratos, qu
```

Instância do assistente

```
TESTE:
Na botânica, quais as classificações da manga?
['botânica, quais classificações manga?']
SGDClassifier
Frutas
0.9890893366337395
BERT:
ML:
Categoria: Frutas
Tema: Manga
Pergunta: Por que a manga escurece?
Resposta: A manga apresenta substâncias chamadas enzimas que, ao entrarem em contato com
```

```
TESTE:
Na botânica, qual a classificação da manga?
['botânica, classificação manga?']
SGDClassifier
Frutas
0.9655609228587035
BERT:
ML:
Categoria: Frutas
Tema: Manga
Pergunta: Qual a classificação botânica da mangueira?
Resposta: A mangueira (Mangifera indica L.) pertence à classe Dicotiledônea e à fa
```

```
TESTE:
Há um buraco negro no centro de nossa galáxia?
['buraco negro centro galáxia?']
SGDClassifier
Animal
0.6883226048853376
SIM cosine_similarity:
Categoria: Graos
Tema: Soja
Pergunta: O que é o centro de origem e o centro de domesticação de uma espécie?
Resposta: O centro de origem ou centro de formação primária de uma espécie é a ár
```

Considerações finais

- Stemming diminuiu em 1% a acurácia do melhor classificador. Correção ortográfica piorou o resultado (ex. Palavras como Levedura);
- Treinamos um modelo **word2vec** (acurácia 0.65), mas a performance não foi boa - poucos dados e baixa variação em perguntas com mesmo sentido;
- No futuro, usaremos predição relacionada ao contexto;
- Será necessário criar uma nova classe ‘outros’ para tratar perguntas que não compõe o conjunto temático.
- Precisamos de mais dados!!!

Obrigado!

wellington.santos@embrapa.br