Trabalho 2 da disciplina de Inteligência Artificial: Estratégias de Aprendizagem em Prompts: Comparação entre Few-Shot e Zero-Shot Learning para Classificação de Intenções Bancárias

Guilherme Hoffmann, João Gabriel Chagas Sbardelotto, Bernardo Wesoly

[1]Escola Politécnica - PUCRS
Porto Alegre – Brasil

Abstract. *Este trabalho apresenta um estudo comparativo entre estratégias de aprendizagem few-shot e zero-shot aplicadas à classificação de intenções bancárias. Utilizando o dataset Banking77 e diferentes Large Language Models (LLMs), avaliamos a eficácia de cada abordagem através das métricas de acurácia, precisão, recall e F1-score. Os resultados demonstram que o few-shot learning supera consistentemente o zero-shot learning, evidenciando a importância de exemplos contextuais para tarefas de classificação especializadas.*

1. Introdução

Com o avanço dos *Large Language Models* (LLMs), novas técnicas de aprendizagem emergiram como alternativas eficientes aos métodos tradicionais de *machine learning*. Entre essas técnicas, destacam-se o *few-shot learning* e o *zero-shot learning*, que permitem que modelos realizem tarefas específicas com poucos ou nenhum exemplo de treinamento, respectivamente.

O ***zero-shot learning*** consiste em instruir o modelo a realizar uma tarefa apenas com descrições textuais, sem fornecer exemplos concretos. Já o ***few-shot learning*** fornece alguns exemplos demonstrativos no *prompt*, permitindo que o modelo aprenda o padrão esperado através de analogia.

A classificação de intenções em sistemas de atendimento bancário representa um desafio relevante para instituições financeiras, pois permite o direcionamento automático de solicitações de clientes para os setores apropriados.

Objetivo: Avaliar quantitativamente a diferença de desempenho entre abordagens *few-shot* e *zero-shot*, além de comparar o comportamento de diferentes LLMs nesta tarefa específica.

2. Solução Desenvolvida

2.1. *Dataset*

Para os experimentos, utilizamos o *dataset* **Banking77**, um conjunto de dados público contendo 13.083 consultas de clientes de serviços bancários, categorizadas em 77 intenções distintas.

Configuração utilizada:

- **Número de classes:** 20 (das 77 disponíveis)
- **Amostras por classe no teste:** 5
- **Total de mensagens de teste:** 100
- **Exemplos *few-shot*:** 20 (1 por classe)

2.2. *Prompt Zero-Shot*

O *prompt zero-shot* instrui o modelo a classificar mensagens sem fornecer exemplos prévios:

```
You are an expert in classifying banking customer support messages.
Your task is to classify the message below into one of the intents
from the provided list.

AVAILABLE INTENTS:
{intents_list}

MESSAGE TO CLASSIFY:
{message}

Respond ONLY with the exact intent name from the list that best
matches the message.
Do not add explanations or additional text.
```

2.3. *Prompt Few-Shot*

O *prompt few-shot* inclui um exemplo por categoria antes da mensagem a ser classificada:

```
You are an expert in classifying banking customer support messages.
Your task is to classify the message below into one of the intents
from the provided list.

EXAMPLES:
{examples}

AVAILABLE INTENTS:
{intents_list}

MESSAGE TO CLASSIFY:
{message}

Respond ONLY with the exact intent name from the list that best
matches the message.
Do not add explanations or additional text.
```

Os exemplos são extraídos do conjunto de treinamento, selecionando uma amostra representativa de cada categoria.

2.4. *System Prompt*

Ambas as abordagens utilizam um *system prompt* comum:

```
You are an assistant specialized in classifying customer support
intents for a banking application.
```

Modelo	Tipo	Parâmetros	Execução
Gemma 2 (2B)	*Open-source* (Google)	2B	Local via Ollama
GPT-4o-mini	Comercial (OpenAI)	N/A	API

Tabela 1. Modelos utilizados nos experimentos

3. Experimentos

3.1. Modelos Utilizados

3.2. Parâmetros dos Modelos

Para garantir reprodutibilidade e consistência:

- ***Temperature*:** 0 (determinístico)
- ***Max tokens*:** 50

3.3. Métricas de Avaliação

- **Acurácia:** Proporção de classificações corretas
- **Precisão (*Weighted/Macro*):** Proporção de verdadeiros positivos entre as predições positivas
- ***Recall* (*Weighted/Macro*):** Proporção de verdadeiros positivos entre os exemplos reais da classe
- **F1-Score (*Weighted/Macro*):** Média harmônica entre precisão e *recall*

3.4. Experimento Adicional

Como experimento adicional, realizamos a comparação entre dois modelos de diferentes capacidades: Gemma 2 (2B), um modelo local e mais leve, e GPT-4o-mini, um modelo comercial mais robusto. Esta comparação permite avaliar o *trade-off* entre custo/privacidade e desempenho.

4. Resultados

4.1. Resultados com Gemma 2 (2B)

Métrica	*Few-Shot*	*Zero-Shot*
Acurácia	87.00%	73.00%
Precisão (*weighted*)	0.9101	0.7648
Recall (*weighted*)	0.8700	0.7300
F1-Score (*weighted*)	0.8711	0.7045
Precisão (*macro*)	0.8274	0.7284
Recall (*macro*)	0.7909	0.6952
F1-Score (*macro*)	0.7919	0.6709

Tabela 2. Resultados com Gemma 2 (2B)

Classes com mais erros (*Few-Shot*): card_about_to_expire (3), balance_not_updated_after_cheque_or_cash_deposit (2), beneficiary_not_allowed (2), card_linking (2), card_payment_not_recognised (2)

Classes com mais erros (*Zero-Shot*): card_arrival (5), card_swallowed (5), card_about_to_expire (4), beneficiary_not_allowed (3), Refund_not_showing_up (2)

Métrica	***Few-Shot***	***Zero-Shot***
Acurácia	97.00%	85.00%
Precisão (*weighted*)	0.9750	0.8888
Recall (*weighted*)	0.9700	0.8500
F1-Score (*weighted*)	0.9697	0.8437
Precisão (*macro*)	0.9750	0.8888
Recall (*macro*)	0.9700	0.8500
F1-Score (*macro*)	0.9697	0.8437

Tabela 3. Resultados com GPT-4o-mini

4.2. Resultados com GPT-4o-mini

Classes com mais erros (*Few-Shot*): balance_not_updated_after_cheque_or_cash_deposit (1), card_delivery_estimate (1), card_payment_fee_charged (1)

Classes com mais erros (*Zero-Shot*): beneficiary_not_allowed (3), card_arrival (3), balance_not_updated_after_bank_transfer (2), card_about_to_expire (2), card_linking (2)

4.3. Análise Comparativa

Os resultados demonstram que o ***few-shot learning*** **apresenta desempenho superior ao** ***zero-shot learning*** em ambos os modelos testados:

- **Gemma 2:** Ganho de 14 pontos percentuais em acurácia (87% vs 73%)
- **GPT-4o-mini:** Ganho de 12 pontos percentuais em acurácia (97% vs 85%)

A comparação entre os modelos revela que o GPT-4o-mini supera o Gemma 2 em todas as métricas, o que é consistente com a diferença de escala e capacidade entre os modelos. O GPT-4o-mini alcançou 97% de acurácia com *few-shot*, errando apenas 3 das 100 mensagens. No entanto, o Gemma 2 apresenta resultados competitivos (87%) considerando suas limitações de tamanho (2B parâmetros) e o fato de ser executado localmente sem custos de API.

4.4. Análise de Erros

As confusões mais frequentes ocorrem entre categorias semanticamente próximas:

Gemma 2 - *Few-Shot*:

- `card_about_to_expire` → `card_delivery_estimate` (3x)

Gemma 2 - *Zero-Shot*:

- `card_arrival` → `card_delivery_estimate` (5x)
- `card_swallowed` → `card_not_working` (5x)

GPT-4o-mini - *Zero-Shot*:

- `beneficiary_not_allowed` → `cancel_transfer` (3x)
- `card_arrival` → `card_delivery_estimate` (3x)

O padrão mais evidente é a confusão entre `card_arrival` e `card_delivery_estimate`, presente em ambos os modelos. Isso ocorre porque ambas as intenções tratam de aspectos relacionados à entrega do cartão, com diferença sutil: uma sobre quando o cartão chegará e outra sobre onde está o cartão. Esse tipo de sobreposição semântica representa o principal desafio para a classificação automática.

5. Conclusão

Este trabalho apresentou uma comparação sistemática entre estratégias de *few-shot* e *zero-shot learning* para a tarefa de classificação de intenções bancárias. Os experimentos demonstraram que:

1. O ***few-shot learning* supera consistentemente o *zero-shot learning***, com ganhos significativos em todas as métricas avaliadas.
2. **Modelos maiores** (GPT-4o-mini) apresentam melhor desempenho, mas **modelos menores** (Gemma 2) podem ser alternativas viáveis para aplicações com restrições de custo ou privacidade.
3. A **qualidade e representatividade dos exemplos** fornecidos no *few-shot learning* são fatores críticos para o desempenho do sistema.

Trabalhos futuros: Investigação de técnicas de *chain-of-thought prompting*, otimização automática de *prompts*, e comparação com abordagens tradicionais de *machine learning* como BERT *fine-tuned*.

Referências

[1] RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence – a Modern Approach**. 4ed. Pearson Education, 2021. 1170p.

[2] GEFFNER, Hector; BONET, Blai. **A Concise Introduction to Models and Methods for Automated Planning**. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers, 2013.

[3] MURPHY, Kevin P. **Machine learning: A probabilistic perspective**. The MIT Press, 2022. 864p.