Teste do *Orca2* de baixo custo para resolução de entidades de produtos eletrônicos

Rodolfo Bolconte Donato (IFPB, Campus João Pessoa), Tiago Brasileiro Araújo (IFPB, Campus Soledade)

E-mails: rodolfo@copin.ufcg.edu.br, tiago.brasileiro@ifpb.edu.br.

Área de conhecimento (Tabela CNPq): 1.03.01.01-1 Computabilidade e Modelos de Computação

Palavras-chave: Resolução de Entidades; Entity Matching; Large Language Models; Base de Dados.

1. Introdução e Justificativa

Na era da informação, caracterizada pela abundância e diversidade de dados computacionais, a organização sistemática dessas informações tornou-se indispensável. Um dos métodos para alcançar essa estruturação é a utilização de referências, que envolvem a indexação de entidades. No campo da Ciência e Engenharia de Dados, essa tarefa é conhecida como Resolução de Entidades (*Entity Matching*, em inglês). Esse processo, essencial para o Processamento de Linguagem Natural (PLN), busca inicialmente identificar e conectar informações, para então determinar se duas ou mais entidades correspondem ao mesmo objeto no mundo real. Fundamental para a limpeza e fusão de dados, tanto em conjuntos únicos quanto distribuídos, a execução precisa e ágil desta tarefa tem aplicações diretas em setores comerciais, científicos e de segurança, embora continue sendo um desafio persistente na integração e limpeza de dados (Zhang; Huan; Joyce, 2024).

Pesquisas recentes exploram a resolução de entidades por meio de abordagens centradas em *tokens*, onde métodos de Aprendizagem Profunda se consolidaram como padrão para essa tarefa. Utilizando exemplos rotulados, os *Pré-trained Language Models (PLMs)* surgem como soluções eficazes, identificando características relevantes das entidades com desempenho satisfatório. No entanto, os *PLMs* tendem a aprender correlações incertas a partir dos dados de treinamento, motivados em parte pelo desequilíbrio na quantidade de dados de classes disponíveis (Akbarian; Mehdi; Davood, 2022).

Os avanços recentes no processamento de linguagem natural viabilizaram o desenvolvimento dos *Large Language Models (LLMs)*, modelos treinados com vastas quantidades de dados textuais, capazes de compreender e gerar texto com nível avançado de naturalidade (Kuang *et al.*, 2024). Com aplicações em diversos domínios, como *chatbots*, assistentes de escrita, mecanismos de busca e sistemas modais, torna-se essencial explorar o potencial dos *LLMs* também na resolução de entidades. A principal vantagem desses modelos sobre os *Pré-trained Language Models* é sua capacidade de executar tarefas sem depender de grandes volumes de exemplos específicos de treinamento, tornando-os uma alternativa promissora para essa atividade (Peeters; Christian, 2023).

Dado o crescente uso de *LLMs* não apenas na comunidade de computação, mas também em diversas outras áreas, torna-se essencial conduzir estudos experimentais para avaliar sua viabilidade na tarefa de Resolução de Entidades. Esses testes devem verificar a precisão dos resultados e o desempenho computacional exigido para a execução dessas tarefas, garantindo que essa abordagem seja de fato adequada para esse contexto.

2. Materiais e métodos

A ideia base do presente estudo é a coleta de um conjunto de dados voltado à atividade de resolução de entidades para que suas informações possam ser enviadas à uma instância de um *LLM*, a fim do modelo ser capaz de dizer se tais informações passadas são referentes a uma mesma entidade ou não. Com as respostas do modelo, é necessária uma validação dos seus erros e acertos, afim de determinar o quão bom o modelo pode ser na realização desta tarefa.

2.1 Conjunto de Dados

Comumente utilizado pela comunidade de computação para a realização de estudos sobre resolução de entidades [(Peeters; Christian, 2023), (Steiner; Ralph; Christian, 2024)], o conjunto de dados *Abt-Buy* foi escolhido na execução do presente trabalho.

Figura 1 – Amostra do Conjunto *Abt-Buy* com id, nome, descrição e preço da entidade.

```
id,nome,descricao,preco
1,bose acoustimass 5 series iii speaker system am53bk,bose acoustimass 5 series iii speaker system
am53bk 2 dual cube speakers with two 2-1/2 ' wide-range drivers in each speaker powerful bass module
with two 5-1/2 ' woofers 200 watts max power black finish,399.0
```

Fonte: Autores do Artigo.

O *Abt-Buy* se trata de um conjunto de dados com informações de produtos eletrônicos, sendo 1081 entidades provenientes do site *abt.com* e 1092 entidades do site *buy.com,* em que cada entidade contém três informações descritivas: nome, descrição e preço do produto. Na Figura 1 é possível visualizar como uma entidade de um produto é descrita no conjunto de dados, com seu identificador do conjunto e demais atributos. Para a resolução de entidades, o conjunto de dados é disposto a partir da associação de duas entidades e seu respectivo rótulo, seja verdadeiro ou falso, sendo assim, a quantidade total de associações com duas entidades que representam o mesmo produto é de 1028 (associação positiva ou verdadeira), enquanto que a quantidade de associações que não representam a mesma entidade é de 8547 (associação negativa ou falsa) (Primpeli; Bizer, 2018).

2.2 Modelo Executado

O modelo escolhido para realizar a tarefa de resolução de entidades foi o *Orca2* com 7 bilhões de parâmetros, que é um modelo desenvolvido pela *Microsoft* com o aprimoramento de habilidades de raciocínio a partir de um versão refinada do *Llama-2*, unicamente para fins de pesquisa afim da própria comunidade realizar avaliações e também melhorias sobre o mesmo (Mitra *et al., 2023*).

Para que fosse possível a criação e execução de uma instância do *Orca2* capaz de realizar a tarefa de resolução de entidades, foi utilizado o *framework Ollama* com instruções de funcionamento específicas, para que sua entrada fosse descrições de duas entidades, e sua saída fosse verdadeiro ou falso de acordo com o seu entendimento caso as duas entidades representem o mesmo produto ou não. Na Figura 2 é possível visualizar o arquivo de criação de uma instância do *Orca2* através do *Ollama,* com descrições do comportamente da instância e também exemplos de entrada e saída que ela deve operar.

Figura 2 – Instruções para a criação de uma instância do *Orca2* através do *Ollama*.

```
FROM orca2

SYSTEM You are a crowdsourcing worker, working on an entity resolution task. You will be given two record descriptions and your task is to identify if the records refer to the same entity or not. You must answer with just one word: True. if the records are referring to the same entity, False. if the records are referring to a different entity.

MESSAGE user record 1: Sony MDREX35LP VB Colorful Headphone with Case - Violet BlueMDREX35LPVB13.540.05Sony7.25 x 2.0 x 1.25 inches record 2: Sony MDR-EX35LP VB EX Style Headphones with Deep Bass Sound Violet BlueMDR-EX35LPVB12.991Sony7.2 x 2.0 x 1.2 inches

MESSAGE assistant True.

MESSAGE user record 1: Sony MDREX35LP VB Colorful Headphone with Case - Violet BlueMDREX35LPVB13.540.05Sony7.25 x 2.0 x 1.25 inches record 2: Sony MDRJ10 LTPNK Clip Style Headphones PinkMDRJ10LTPNK9.061Sony7.2 x 4.0 x 1.5 inches

MESSAGE assistant False.
```

Fonte: Autores do Artigo.

2.3 Métricas para Análise

Visando uma validação coerente para a atividade de resolução de entidades e por se tratar de uma classificação binária de informações, são utilizadas as seguintes métricas (Donato, 2024): 1) *Precision:* a proporção de previsões positivas verdadeiras para o total de previsões positivas; 2) *Recall:* avalia as amostras positivas previstas corretamente em relação à quantidade de amostras positivas originais do conjunto de dados; e 3) *F1-Score:* a média harmônica de *Precision* e *Recall*, dando um peso maior para os valores baixos. Com tais métricas, será possível mensurar o desempenho do *Orca2* na realização da resolução de entidades com o conjunto de dados *Abt-Buy*, e então definir – apenas para os aspectos do presente trabalho – se tal modelo é viável ou não para a realização da tarefa.

3. Resultados e Discussão

No total são realizadas 3 execuções com quantidades diferentes de dados, sendo: a 1ª execução com 1710 amostras negativas e 206 positivas; a 2ª com 8547 negativas e 1028 positivas; e a 3ª execução com 1028 amostras positivas e o mesmo número para as amostras negativas, esta sendo uma execução balanceada de classes.

Em todas as execuções o *Orca2* consegue resultados satisfatórios para as três métricas utilizadas, quando levando em consideração o desempenho do modelo para ambas as classificações possíveis, porém ao analisar as métricas para cada uma das classes individualmente, o cenário não é satisfatório.

Na primeira execução com 1916 amostras, o modelo consegue atingir valores próximos dos 90% para as três métricas ao analisar todas as associações. Para as associações negativas, os valores são ainda melhores com 98% de *Recall.* Porém, se tratando das associações positivas, isto é, quando as duas entidades enviadas são referentes ao mesmo produto, os valores não são satisfatórios, com o *F1-Score* atingindo apenas 41%. Na Tabela 1 é possível verificar todos os valores da primeira execução do modelo.

Tabela 1 – Valores métricos da primeira execução com 1916 associações de entidades.

Associações	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Todas (1916)	89	91	89
Negativas (1710)	92	98	95
Positivas (206)	62	31	41

O resultado desta primeira execução já mostra o que deve ser esperado para as execuções seguintes, que devido ao desbalanceamento da quantidade de amostras por classe, 1710 amostras negativas para apenas 206 amostras positivas, pode induzir o modelo para uma maior quantidade de falsos negativos em seus resultados, assim como também a própria capacidade do modelo na identificação de entidades através de nome, descrição e preço das mesmas.

Na segunda execução com 9575 amostras, sendo 8547 negativas para 1028 amostras positivas, o modelo atinge valores próximos e até iguais em relação a primeira execução para todas as associações e as associações negativas, porém para as associações positivas o resultado é menor, como mostrado na Tabela 2. Levando em consideração as 1028 amostras positivas, o *Orca2* consegue apenas 36% de *F1-Score,* evidenciando que quanto mais amostras iguais, menos o modelo é capaz de perceber tais semelhanças.

Tabela 2 – Valores métricos da segunda execução com 9575 associações de entidades.

Associações	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Todas (9575)	88	90	88
Negativas (8547)	92	98	95
Positivas (1028)	58	26	36

Fazendo uma comparação das duas primeiras execuções, pode ser dito que a diminuição dos valores métricos para a classe positiva não era algo esperado, uma vez que a proporção da quantidade de dados se manteve a mesma nas duas execuções, sendo cerca de 89,25% dos dados negativos e 10,75% dos dados positivos.

Por fim, para a terceira execução foi idealizado um cenário com dados balanceados, para verificar se com quantidades iguais nas duas classes o modelo geraria resultados melhores, uma vez que não há um desbalanceamento de classes que possa diminuir os valores métricos para as amostras da classe positiva. Infelizmente, não só os resultados continuam não promissores para as amostras positivas, como há uma piora ao levar em conta todas as amostras e somente as amostras negativas nas métricas. Na Tabela 3 é possível visualizar que o modelo obtém apenas 56% de *F1-Score* para todas as associações e 72% para as associações negativas, estes valores sendo 33 e 23 pontos percentuais menores que o *F1-Score* da primeira execução, respectivamente. Para as associações positivas, o *F1-Score* aumentou em relação a segunda execução porém diminuiu em relação a primeira, evidenciando novamente o baixo desempenho classificatório do modelo para amotras positivas.

Tabela 3 – Valores métricos da terceira execução com 2056 associações de entidades.

Associações	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Todas (2056)	74	62	56
Negativas (1028)	57	98	72
Positivas (1028)	92	25	40

5. Considerações finais

Com os resultados obtidos no presente estudo, é evidente que o *Orca2* não se mostra como um modelo adequado para realizar atividade de resolução de entidades, porém, esta afirmação cabe somente aos dados utilizados do conjunto *Abt-Buy* e também aos atributos padrão utilizados em sua execução, uma vez que não é possível generalizar os seus resultados para outros tipos de dados e parâmetros.

Apesar do desempenho não favorável, é interessante levar em consideração os acertos que o modelo obteve visando a realização de estudos dos casos positivos acertados, a fim de verificar toda a lógica por trás do resultado e confirmar se há abertura para novas execuções e melhora do modelo ou não. Outra possibilidade de investigação seria a utilização do *Orca2* apenas para a execução em amostras que gerem conflitos em outros modelos, tornando-o como o modelo final de classificação.

Referências

AKBARIAN RASTAGHI, MEHDI; EHSAN KAMALLOO; DAVOOD RAFIEI. **Probing the robustness of pre-trained language models for entity matching.** Proceedings of the 31st ACM International Conference on Information & Knowledge Management. (2022).

PEETERS, RALPH; CHRISTIAN BIZER. **Entity matching using large language models.** arXiv preprint arXiv:2310.11244 (2023).

DONATO, RODOLFO BOLCONTE. **Análise de transformações metamórficas em conjunto de dados para garantia de fairness em modelos de classificação de informações.** (2024).

KUANG, WEIRUI, et al. **Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning.** Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. (2024).

ZHANG, JING; HUAN SUN; JOYCE C. HO. **EMBA: Entity Matching using Multi-Task Learning of BERT with Attention-over-Attention.** EDBT. (2024).

PRIMPELI, A.; BIZER, C. **Profiling entity matching benchmark tasks.** In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 3101-3108). Disponível em https://dl.acm.org/doi/abs/10.1145/3340531.3412781. (2020).

MITRA, A.; DEL CORRO, L.; MAHAJAN, S.; CODAS, A.; SIMOES, C.; AGARWAL, S.; AWADALLAH, A. **Orca 2: Teaching small language models how to reason**. arXiv preprint arXiv:2311.11045. Disponível em: https://arxiv.org/abs/2311.11045. (2023).

STEINER, AARON; RALPH PEETERS; CHRISTIAN BIZER. **Fine-tuning Large Language Models for Entity Matching.** arXiv preprint arXiv:2409.08185. (2024).