# AspectRAG: Uma Arquitetura de Recuperação e Geração para Análise de Sentimentos Baseada em Aspectos

**Erick R. Ribeiro[1], André Carvalho[1], Rhedson Esashika[2]**

[1]Instituto de Computação, Universidade Federal do Amazonas (UFAM), Manaus, AM, Brasil
[2]Universidade Estadual do Amazonas (UEA), Manaus, AM, Brasil
{erick, andre}@icomp.ufam.edu.br, rffe.mee25@uea.edu.br

## Resumo

Neste trabalho propomos o AspectRAG, uma arquitetura de Recuperação e Geração para Extração de Triplas de Sentimento de Aspectos (ASTE) em português que opera sem treinamento supervisionado. O método extrai aspectos com um LLM, codifica-os como vetores densos e usa apenas esses vetores para recuperar evidências altamente específicas por meio de busca aproximada e fusão de rankings. As evidências recuperadas compõem o contexto do modelo gerador, que produz as triplas finais. Nos datasets ReLi e ReHol, o AspectRAG obtém até 93,47% em ATE, 80,68% em OTE e 69,83% em ASTE, superando modelos supervisionados como OTE-MTL, CMLA-MTL e BOTE, o estado da arte em Português. Nosso estudo de ablação evidencia que a recuperação semântica guiada por aspectos é o principal fator responsável pelos ganhos observados, enquanto a quantidade de parâmetros do LLM tem impacto secundário. Os resultados mostram que a arquitetura AspectRAG é uma solução eficiente e competitiva mesmo sem fine-tuning, apoiando-se apenas em recuperação vetorial e inferência contextualizada.

## 1 Introdução

Com o advento da Web 2.0, consumidores passaram a compartilhar suas experiências sobre produtos, serviços e destinos em plataformas online (Yoo e Gretzel, 2008). Desde então, esses textos opinativos constituem uma valiosa fonte de informação para empresas e pesquisadores, permitindo compreender percepções, medir níveis de satisfação e orientar decisões estratégicas (Chevalier e Mayzlin, 2006; Cortis e Davis, 2021). Neste contexto, a Análise de Sentimentos, também conhecida como mineração de opinião, emergiu como um campo de estudos da área de Processamento de Linguagem Natural, voltada a identificar automaticamente a polaridade a partir da linguagem escrita (Liu e Zhang, 2012; Pontiki et al., 2016). Contudo, as abordagens tradicionais de análise de sentimentos classificavam os textos inteiros como positivos, negativos ou neutros, sendo limitados quando uma mesma sentença expressa opiniões divergentes sobre diferentes aspectos do assunto abordado (Zhang et al., 2022). Essa limitação impulsionou o desenvolvimento da Análise de Sentimentos Baseada em Aspectos (ABSA, *aspect based sentiment analysis*). A ABSA recebe textos opinativos, identifica de forma granular os aspectos mencionados, as opiniões associadas e suas respectivas polaridades (Thet et al., 2010).

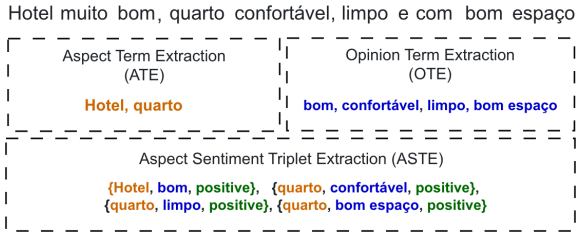



Figura 1: Exemplo das subtarefas de ABSA.

A formulação completa de ABSA objetiva capturar quatro elementos que compõem as opiniões expressas em textos opinativos: termos de aspectos, categorias de aspecto, os termos de opinião e a polaridade associada a cada relação aspecto-opinião (Pereira, 2020). Como muitos conjuntos de dados não incluem essa quádrupla completa, a tarefa de ABSA costuma ser decomposta em subtarefas com objetivos e complexidade distintas, como ilustrado na Figura 1. Entre elas, destacam-se: a Extração de Termos de Aspecto (ATE), que identifica os aspectos mencionados (Li et al., 2018; Chen e Qian, 2020); a Extração de Termos de Opinião (OTE), que detecta as expressões opinativas (Barros e De Bona, 2021; Chakraborty, 2024); e a Extração de Triplas de Sentimento de Aspectos (ASTE), que reúne aspecto, opinião e polaridade em triplas (Jian et al., 2021; Naglik et al., 2024). A ASTE oferece uma análise mais detalhada ao



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 194–202

explicitar essas relações, mas também traz desafios significativos, pois exige capturar relações sintáticas e semânticas complexas.

Apesar dos avanços recentes nesta seara, aplicações para o português permanecem pouco exploradas, e as soluções atuais dependem majoritariamente de métodos supervisionados ou baseados em regras, que requerem dados anotados e alto custo computacional (Câmara e Melo, 2022; Pereira et al., 2024; Machado et al., 2021b). Por outro lado, arquiteturas de Geração Aumentada por Recuperação (RAG) permitem incorporar conhecimento externo sem treinamento, ampliando a aplicabilidade de LLMs em cenários de baixo recurso (Lewis et al., 2020b; Karpukhin et al., 2020; Izacard e Grave, 2020; Guu et al., 2020).

Embora já existam trabalhos que exploram recuperação de vizinhos semânticos para tarefas relacionadas a ABSA/ASTE, o uso de uma arquitetura RAG totalmente integrada e direcionada às subtarefas de ASTE ainda é pouco explorado, especialmente para o português (Yu et al., 2023; Wang e et al., 2024). Diante desse cenário, este trabalho propõe o AspectRAG, uma nova arquitetura de RAG voltada para a tarefa de ABSA, com ênfase nas subtarefas ATE, OTE e ASTE. O AspectRAG integra um mecanismo de recuperação semântica orientado a aspectos com um modelo gerativo especializado na identificação de termos de aspecto, opinião e polaridade. Os resultados experimentais demonstram que o AspectRAG, com GPT-4o-mini e Qwen-3-8B, supera consistentemente os baselines em todas as subtarefas e datasets, evidenciando que a principal contribuição do método reside na estrutura de recuperação orientada a aspectos, e não está limitado a um único modelo gerador. As principais contribuições deste trabalho podem ser resumidas da seguinte forma:

- Propomos o AspectRAG, a primeira arquitetura RAG adaptada às subtarefas de ABSA em português;
- Mostramos que a recuperação guiada por aspectos melhora a qualidade da extração de termos e triplas;
- Demonstramos ganhos consistentes sobre modelos supervisionados e prompting em múltiplos datasets;

## 2 Trabalhos Relacionados

A tarefa de ABSA tem evoluído de abordagens supervisionadas tradicionais para métodos baseados em LLMs e arquiteturas híbridas de recuperação e geração. Em português, a maioria dos trabalhos concentra-se na detecção isolada de aspectos ou polaridades, utilizando métodos baseados em regras (Câmara e Melo, 2022), abordagens híbridas (Machado et al., 2021a) ou técnicas supervisionadas aplicadas a domínios específicos (Pereira et al., 2024). Mais recentemente, (Seno et al., 2024) investigam o uso de LLMs para identificar aspectos e polaridades em comentários sobre debates políticos, avaliando modelos como o GPT em diferentes configurações de *prompting*. Esses estudos, entretanto, não contemplam a formulação completa da tarefa ASTE, que exige a identificação simultânea de aspectos, opiniões e polaridades, nem exploram mecanismos de recuperação de evidências ou arquiteturas híbridas de recuperação e geração.

Dentro da literatura supervisionada de ASTE, modelos como OTE-MTL e CMLA-MTL realizaram a coextração de aspectos e opiniões por meio de arquiteturas multitarefa (Wang et al., 2017; Zhang et al., 2020). Para o português, o estado da arte é o BOTE (Barros e De Bona, 2021), que combina BERTimbau, BiLSTM, atenção gráfica e um biaffine scorer para prever dependências entre aspectos e opiniões.

Com a ascensão dos LLMs, técnicas de *prompting* como Zero-Shot e Few-Shot passaram a ser exploradas como alternativa sem treinamento supervisionado. Apesar disso, tais abordagens ainda enfrentam dificuldades para segmentar precisamente aspectos e opiniões e para associar corretamente polaridades, especialmente em textos longos ou heterogêneos. Nesse cenário, arquiteturas de Recuperação Aumentada por Geração (RAG) surgem como uma alternativa promissora, pois combinam recuperação de evidências externas com geração contextualizada (Lewis et al., 2020a; Guu et al., 2020; Izacard e Grave, 2020).

Embora arquiteturas RAG já tenham sido aplicadas em diferentes tarefas de NLP (Yu et al., 2023; Wang e et al., 2024), esses trabalhos operam principalmente em inglês e não utilizam recuperação explícita orientada a aspectos. O método proposto neste trabalho, o AspectRAG, introduz uma nova estratégia de recuperação orientada exclusivamente por aspectos extraídos da sentença. Essa decisão de projeto simplifica o processo de recuperação e

permite recuperar evidências altamente específicas para cada relação aspecto-opinião, o que se mostra particularmente eficaz para a tarefa de ASTE em português.

## 3 Método proposto

### 3.1 Definição do Problema

Segundo (Barros e De Bona, 2021), dada uma sentença $S = \{w_1, w_2, \ldots, w_n\}$ composta por $n$ palavras, o objetivo da tarefa ASTE é extrair um conjunto de triplas $T = \{(a, o, p)\}_{m=1}^{|T|}$ presentes em $S$. Nessa formulação, $a$, $o$ e $p$ representam, respectivamente, um termo de aspecto, um termo de opinião e uma polaridade de sentimento. Os termos $a_m$ e $o_m$ são segmentos textuais, sem alteração, da sentença $S$, e podem ser representados por suas posições inicial e final $(inicio_m, fim_m)$ na sentença. Já a polaridade $p_m$ pertence ao conjunto $\{\text{Positivo}, \text{Negativo}, \text{Neutro}\}$.

A tarefa ASTE envolve relações complexas entre aspectos e opiniões, podendo ser classificadas em três tipos: (i) *1:1*, quando um aspecto está associado a uma única opinião; (ii) *1:M*, quando um aspecto está vinculado a múltiplas opiniões; e (iii) *M:N*, quando múltiplos aspectos estão associados a múltiplas opiniões simultaneamente. No contexto do AspectRAG, essa mesma formulação é mantida, porém o modelo se destaca por realizar a extração das triplas de forma hierárquica em duas etapas complementares. Primeiro, é realizada a detecção dos aspectos via LLM. Posteriormente, recuperação e geração das triplas, como uma arquitetura de RAG tradicional.

Essa estrutura em duas etapas nos permite capturar de maneira mais precisa e focada os aspectos, que servem para identificar os melhores exemplos para cada tipo de relação aspecto/opinião, desde as mais simples até as mais complexas. Tal característica faz com que o modelo apresente desempenho consistentemente superior aos baselines, uma vez que a etapa inicial de extração de aspectos direciona a recuperação semântica para contextos mais relevantes, potencializando a capacidade do gerador em produzir triplas mais coerentes e completas.

### 3.2 Arquitetura Proposta

A arquitetura completa do *AspectRAG* é ilustrada na Figura 2. Ela é composta por dois módulos principais: Recuperação e Geração. O primeiro identifica os aspectos na sentença e recupera exemplos semanticamente similares. O segundo utiliza essas evidências para produzir as triplas.

O fluxo de processamento pode ser descrito pelas seguintes etapas:

1. **Entrada**: o sistema recebe uma sentença opinativa em linguagem natural. Esse texto é a unidade básica de análise, podendo conter múltiplos aspectos e opiniões.

2. **Extração de Aspectos**: um modelo de linguagem (LLM) identifica automaticamente os termos de aspecto presentes na sentença. Essa etapa é realizada via *prompt engineering* e exemplos *few-shot*, retornando o conjunto $A = \{a_1, a_2, \ldots, a_K\}$.

3. **Codificação de Aspectos**: cada aspecto $a_k$ é transformado em um vetor denso por meio de modelos de geração de *embeddings*. Essa codificação preserva propriedades semânticas, permitindo medir similaridade entre aspectos de diferentes sentenças.

4. **Indexação Vetorial**: os vetores de aspectos são armazenados em uma base vetorial, que permite consultas eficientes via busca de vizinhos aproximados. Cada aspecto é mapeado a exemplos contextualmente semelhantes.

5. **Recuperação de Documentos**: para cada aspecto, o sistema recupera os $N$ documentos mais similares ao vetor de consulta. O resultado é um conjunto de contextos relevantes $C = \{c_1, c_2, \ldots, c_N\}$, contendo exemplos anotados de pares aspecto–opinião–polaridade. [1]

6. **Fusão de Rankings**: quando múltiplos aspectos são detectados em uma mesma sentença, as listas de documentos recuperadas são combinadas por meio da técnica de Reciprocal Rank Fusion. Essa fusão prioriza documentos que aparecem nas primeiras posições de múltiplas listas de recuperação, garantindo diversidade e relevância contextual.

7. **Prompt + Geração de Triplas**: o texto de entrada é concatenado com os exemplos recuperados, formando um *prompt* contextualizado

[1] O termo documento refere-se a uma sentença anotada presente no conjunto de treinamento. Cada documento contém exemplos de triplas (aspecto, opinião, polaridade) previamente anotadas, que são utilizadas como evidência contextual para orientar a geração das novas triplas.

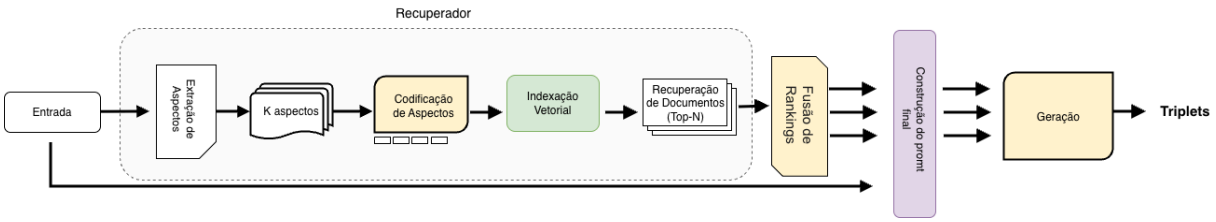



Figura 2: Arquitetura geral do AspectRAG. O módulo de recuperação agrega contextos relevantes a partir dos aspectos extraídos, enquanto o gerador produz as triplas de sentimento.

que alimenta o modelo gerador. O gerador, implementado com um modelo de linguagem (por exemplo, GPT-4o-mini), produz as triplas $(a, o, p)$ a partir da instrução e dos exemplos.

### 3.3 Definição da Recuperação

A recuperação de evidências semânticas no AspectRAG é formalizada como um problema de busca em espaço vetorial, no qual apenas os aspectos extraídos da sentença são utilizados para a construção das consultas. Em contraste com métodos convencionais, o vetor de embeddings da sentença $s$ não participa diretamente da busca, sendo empregado apenas na etapa de composição do *prompt* gerativo.

Para cada aspecto $a_k$ identificado na sentença, é gerado um vetor de consulta definido como:

$$q_k = \text{embed}(a_k)$$

Esse vetor é então comparado a todos os vetores $d_i$ armazenados no índice vetorial, calculando-se a função de similaridade $\text{sim}(q_k, d_i)$ por meio do cosseno entre os vetores. O conjunto dos $N$ documentos mais similares para cada aspecto é selecionado conforme:

$$C_k = \text{Top-}N(\text{sim}(q_k, D))$$

Cada aspecto, portanto, gera sua própria lista de recuperação, capturando evidências específicas e contextualmente relevantes a ele. Esse comportamento assegura que o AspectRAG selecione exemplos mais apropriados para diferentes tipos de relações entre aspectos e opiniões, desde casos simples até os mais complexos.

As listas associadas a múltiplos aspectos são combinadas por meio da técnica de *Reciprocal Rank Fusion (RRF)*, expressa por:

$$RRF(d) = \sum_{a_k \in A} \frac{1}{k + \text{rank}_{a_k}(d)}$$

Onde $\text{rank}_{a_k}(d)$ indica a posição do documento $d$ na lista de resultados obtida para o aspecto $a_k$. Essa fusão garante que documentos relevantes para diferentes aspectos da mesma sentença recebam maior pontuação agregada, resultando em um conjunto final de contextos semanticamente ricos. Finalmente, o texto original da sentença $s$ é reintroduzido na etapa de composição do *prompt*, sendo concatenado aos exemplos recuperados para orientar o modelo gerativo na produção das triplas. A próxima seção descreve os experimentos realizados para avaliar o desempenho do AspectRAG e validar as contribuições discutidas até aqui.

## 4 Metodologia de Avaliação

### 4.1 Datasets

Nós conduzimos experimentos em dois conjuntos de dados em Português, preparados por (Barros e De Bona, 2021), e disponíveis no IEEE Dataport[2]. Os datasets ReLi e ReHol, ambos contêm anotações manuais de aspectos, opiniões e polaridades, adequados para avaliação das subtarefas de ATE, OTE e ASTE. O primeiro conjunto, denominado *ReLi*, reúne resenhas literárias de 14 obras distintas, caracterizadas por linguagem subjetiva, estruturas sintáticas heterogêneas e grande variação estilística, tornando-o um domínio intrinsecamente desafiador para sistemas de extração.

O segundo conjunto, *ReHol*, contém avaliações de hotéis coletadas do TripAdvisor e apresenta maior objetividade, menor variação sintática e maior ocorrência de múltiplos aspectos por sentença, o que o torna particularmente relevante para avaliar a recuperação orientada por aspectos empregada no *AspectRAG*. A Tabela 1 sintetiza as principais estatísticas desses dois corpora, incluindo a distribuição de triplas anotadas, polaridades e características estruturais das sentenças.

[2]Os datasets podem ser acessados em: https://dx.doi.org/10.21227/0ej1-br13.

Tabela 1: Estatísticas descritivas dos datasets ReLi e ReHol.

| Nomenclatura | ReHol | ReLi |
|---|---|---|
| Domínio | Hotel | Livro |
| # Triplas | 3199 | 1448 |
| # Triplas positivas | 1846 | 1170 |
| # Triplas negativas | 1242 | 278 |
| # Triplas neutras | 111 | 0 |
| # Sentenças não sobrepostas | 941 | 767 |
| # Sentenças sobrepostas | 573 | 242 |
| Média de palavras por sentença | 16 | 22 |
| Tamanho mínimo da sentença (palavras) | 8 | 2 |
| Tamanho máximo da sentença (palavras) | 38 | 193 |

### 4.2 Baselines

Os baselines estão organizados em dois grupos: (i) métodos supervisionados que requerem treinamento de modelos e representam trabalhos publicados relevantes na literatura de ASTE; e (ii) técnicas baseadas em prompting e recuperação, que não exigem treinamento convencional e incluem contribuições originais deste artigo adaptadas às subtarefas ATE, OTE e ASTE.

Os métodos supervisionados que requerem treinamento de modelos são:

- **OTE-MTL** (Zhang et al., 2020) – Modelo que codifica sentenças com GloVe + BiLSTM, seguido por camadas lineares e não lineares para identificar termos de aspecto e opinião e biaffine scorer para determinar dependências de polaridade entre aspectos e opiniões.
- **CMLA-MTL** (Wang et al., 2017) – Modelo baseado em múltiplas camadas de atenção acopladas (Coupled Multi-Layer Attentions), originalmente proposto para coextração de aspectos e opiniões, e posteriormente adaptado para a tarefa ASTE.
- **BOTE** (Barros e De Bona, 2021) – Modelo que combina vetores contextualizados, com BERTimbau, uma camada BiLSTM para redução de dimensionalidade, uma camada de atenção gráfica (GAL) para incorporar relações sintáticas e biaffine scoring para prever dependências aspecto–opinião.

Já as técnicas baseadas em prompting e recuperação são:

- **Zero Shot** – Técnica de engenharia de prompt sem exemplos, na qual o modelo recebe apenas instruções específicas adaptadas para a tarefa ASTE.
- **Few Shot** – Técnica de prompting com exemplos fixos no prompt, adaptada para ASTE adicionando algumas triplas ilustrativas como demonstração.
- **Review RAG** – Pipeline tradicional de RAG, adaptado ao contexto da tarefa de ASTE, no qual o conjunto de treino é indexado no formato (review, triplas). Note que não há extração prévia de aspectos, ou seja, o texto completo da review é utilizado diretamente para recuperar evidências relevantes, diferentemente do AspectRAG.

Os três primeiros métodos representam trabalhos supervisionados publicados, com arquiteturas profundas completas que dependem de treinamento extenso. Os três últimos são técnicas baseadas em prompting ou recuperação adaptadas para este estudo, e não requerem qualquer treinamento supervisionado. A proposta *AspectRAG*, da mesma forma, não necessita de treinamento de modelos, destacando-se pela capacidade de alcançar desempenho competitivo mesmo em cenários com escassez de dados anotados.

### 4.3 Configuração dos experimentos

Os experimentos foram conduzidos utilizando três modelos de linguagem de médio porte: GPT-4o-mini, Qwen-3-4B e Qwen-3-8B[3]. Esses LLMs foram empregados em todas as configurações avaliadas, permitindo comparar consistentemente o comportamento das diferentes técnicas de prompting e de recuperação. Como nossa abordagem e os baselines não supervisionados não envolvem treinamento, nenhum ajuste de parâmetros, fine-tuning ou otimização adicional foi aplicado.

Nas configurações Zero Shot, cada LLM recebeu apenas um prompt instrucional adaptado para as subtarefas ATE, OTE e ASTE, sem exemplos adicionais. Já nas configurações Few Shot, incluímos no prompt um conjunto fixo de $k$ exemplos ilustrativos, adicionados de forma estática, permitindo avaliar o impacto do in-context learning. O valor de $k$ variou conforme o experimento, de modo a mensurar o efeito direto da quantidade de demonstrações na extração das triplas.

[3]Os modelos escolhidos representarem tanto alternativas proprietárias quanto modelos abertos com diferentes capacidades, permitindo avaliar o comportamento da arquitetura proposta em cenários com distintos níveis de capacidade do modelo gerador.

Nas configurações ReviewRAG e AspectRAG, o parâmetro $k$ corresponde ao número de exemplos recuperados dinamicamente pelo módulo de busca. No ReviewRAG, cada consulta utiliza o texto completo da review original como entrada para recuperação, enquanto no AspectRAG a consulta é construída a partir dos aspectos previamente extraídos, possibilitando uma recuperação mais direcionada às relações aspecto–opinião. Em ambos os casos, o conjunto recuperado é incorporado ao prompt do LLM como evidência adicional.

### 4.4 Métricas de Avaliação

A avaliação dos métodos foi conduzida seguindo as mesmas diretrizes adotadas nos trabalhos anteriores sobre ASTE. Utilizamos as métricas clássicas de *precision*, *recall* e *F1-score*, calculadas individualmente para cada subtarefa: ATE, OTE e ASTE. Assim como reportado em (Barros e De Bona, 2021), uma extração é considerada correta apenas se todos os elementos relevantes forem identificados exatamente.

Para a subtarefa ATE, um termo de aspecto é contado como verdadeiro positivo somente quando suas fronteiras textuais (início e fim) correspondem exatamente às anotações de referência. O mesmo critério é aplicado à subtarefa OTE, exigindo correspondência exata das fronteiras dos termos de opinião.

Na subtarefa ASTE, uma tripla é considerada correta apenas quando os três componentes são identificados simultaneamente: o termo de aspecto, o termo de opinião associado, e a polaridade correta da relação. Assim, erros em qualquer um desses elementos, seja na segmentação textual ou na polaridade, tornam a tripla inteira incorreta.

O conjunto de treino e teste foi dividido de forma idêntica ao trabalho de (Barros e De Bona, 2021). Com uso da técnica de validação cruzada aninhada de 5×2, garantindo uma comparação direta entre abordagens supervisionadas e métodos baseados em LLMs.

## 5 Resultados e Análises

### 5.1 Variação do K na Recuperação Semântica

A quantidade de exemplos recuperados ($K$) controla a diversidade contextual inserida no prompt do *AspectRAG*. Avaliamos $K = \{2, 4, 6, 8, 10\}$ utilizando *Qwen-3-8B* como modelo gerador principal. A Figura 3 apresenta a evolução do *F1-Score* em ASTE para cada valor de $K$, incluindo como linha de referência o desempenho supervisionado do BOTE.

A análise dos resultados mostra que valores muito baixos de $K$ reduzem o desempenho em ASTE, pois limitam a variedade de evidências disponíveis para o modelo. No ReLi, observa-se um crescimento consistente do F1 à medida que $K$ aumenta, com melhor resultado em $K = 8$. No ReHol, o ganho é mais gradual e se estende até $K = 10$, o que indica que domínios com maior densidade de aspectos por sentença se beneficiam de conjuntos mais amplos de exemplos recuperados.

De forma geral, o intervalo entre $6 \leq K \leq 10$ oferece o melhor equilíbrio entre diversidade e relevância das evidências, evitando tanto a falta de contexto quanto o excesso de informação pouco útil. Os resultados também mostram que os modelos Qwen-3-8B e GPT-4o-mini apresentam comportamento semelhante em resposta ao aumento de $K$, o que sugere que a qualidade da recuperação exerce influência mais significativa no desempenho do que o tamanho do modelo gerador.

Como $K = 10$ apresenta o melhor desempenho no ReHol e está entre os valores de maior desempenho no ReLi, esse valor foi adotado como configuração padrão para os experimentos subsequentes. Essa escolha garante que a análise dos componentes do *AspectRAG* seja conduzida em um cenário favorável, no qual a recuperação fornece variedade suficiente de evidências para apoiar a geração de triplas.

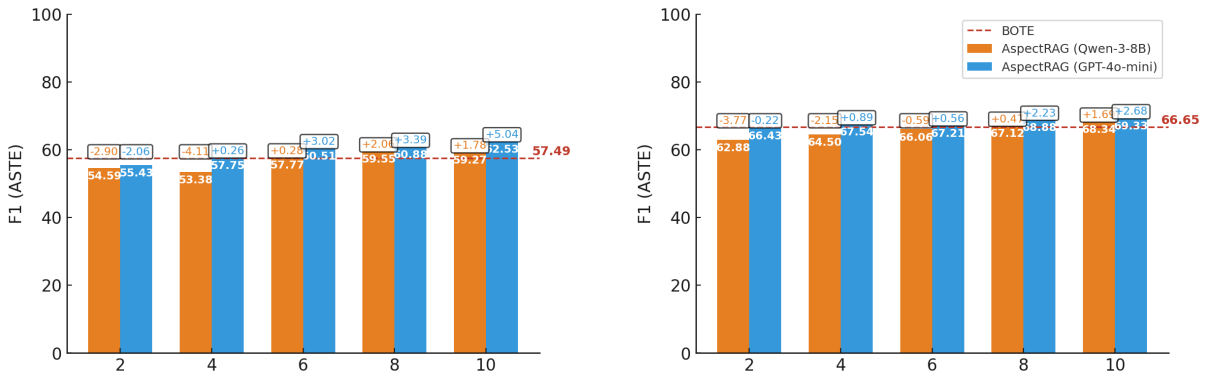
### 5.2 Comparação com Baselines

A Tabela 2 apresenta a comparação entre os baselines supervisionados (OTE-MTL, CMLA-MTL e BOTE) e as variantes do *AspectRAG* com Qwen3-4B e Qwen3-8B. Entre os modelos supervisionados, o BOTE é o mais forte, alcançando 57.49% em ASTE no ReLi e 66.65% no ReHol.

Tabela 2: Comparação com baselines supervisionados usando F1-Score.

| Modelo | ReLi | | | ReHol | | |
|---|---|---|---|---|---|---|
| | ATE | OTE | ASTE | ATE | OTE | ASTE |
| OTE-MTL | 71.48 | 68.61 | 49.92 | 78.70 | 72.93 | 60.91 |
| CMLA-MTL | 72.14 | 62.16 | 40.61 | 74.31 | 69.45 | 47.03 |
| BOTE | 79.49 | 70.52 | 57.49 | 82.20 | 81.10 | 66.65 |
| AspectRAG / GPT4o-mini | 90.89 | **72.40** | 62.53 | 87.48 | **79.90** | **69.83** |
| AspectRAG / Qwen3-4B | 87.66 | 67.36 | 56.52 | 87.23 | 75.30 | 63.13 |
| AspectRAG / Qwen3-8B | **93.47** | 69.96 | **59.27** | **88.38** | 80.68 | 68.33 |

As variantes do *AspectRAG* superam esse desempenho, mesmo sem treinamento supervisionado.




(a) ReLi (b) ReHol

Figura 3: Variação do F1-Score na tarefa ASTE em função de $K$ nos datasets ReLi (a) e ReHol (b). As barras representam os resultados do AspectRAG com diferentes LLMs e a linha tracejada indica o desempenho supervisionado do BOTE.

O modelo com GPT-4o-mini atinge 62.53% em ASTE no ReLi e 69.83% no ReHol, o que representa ganhos de +5.04 e +3.18 pontos percentuais em relação ao BOTE. Embora esse modelo seja proprietário e não tenha seu tamanho divulgado pela OpenAI, ele serve como referência prática de desempenho para sistemas RAG com LLMs comerciais.

Os modelos abertos também apresentam resultados competitivos. O *AspectRAG* com Qwen-3-4B se aproxima dos baselines, obtendo 56.52% em ASTE no ReLi (diferença de –0.97 p.p.) e 63.13% no ReHol (diferença de –3.52 p.p.). Com o aumento de capacidade, o Qwen-3-8B passa a superar todos os baselines em ambos os datasets, alcançando 59.27% no ReLi (+1.78 p.p.) e 68.33% no ReHol (+1.68 p.p.).

Esses resultados mostram que a arquitetura *AspectRAG* permite que diferentes LLMs alcancem desempenho comparável ou superior aos baselines, reforçando o papel central da recuperação orientada a aspectos na construção das triplas de sentimento, mesmo sem qualquer etapa de treinamento específico.

### 5.3 Estudo de Ablação

A Tabela 3 compara variantes do *AspectRAG* com estratégias sem recuperação (Zero Shot e Few Shot) e com o ReviewRAG. Os resultados permitem quantificar o impacto da recuperação orientada a aspectos nas subtarefas ATE, OTE e ASTE.

De forma geral, observa-se que ATE é sempre a subtarefa mais fácil, com valores superiores aos de OTE e ASTE em todos os cenários. Nos modelos sem recuperação, esse efeito é particularmente

Tabela 3: Resultados da ablação, nas subtarefas ATE, OTE e ASTE, nos datasets ReLi e ReHol.

| Modelo / Variante | ReLi | | | ReHol | | |
|---|---|---|---|---|---|---|
| | ATE | OTE | ASTE | ATE | OTE | ASTE |
| Zero Shot / GPT-4o-mini | **73.45** | **67.37** | **55.75** | **79.79** | **72.91** | **60.08** |
| Zero Shot / Qwen-3-4B | 42.20 | 47.30 | 23.30 | 75.04 | 56.34 | 45.75 |
| Zero Shot / Qwen-3-8B | 46.89 | 50.11 | 27.19 | 77.66 | 58.25 | 47.60 |
| Few Shot / GPT-4o-mini | **77.64** | **64.07** | **57.28** | **84.45** | **71.89** | **63.28** |
| Few Shot / Qwen-3-4B | 72.69 | 16.87 | 13.25 | 83.42 | 64.40 | 54.67 |
| Few Shot / Qwen-3-8B | 74.44 | 48.87 | 39.38 | 88.64 | 64.18 | 56.03 |
| ReviewRAG / GPT-4o-mini | **84.44** | **71.37** | **59.87** | **82.83** | **80.61** | **67.57** |
| ReviewRAG / Qwen-3-4B | 79.85 | 67.33 | 54.82 | 81.59 | 74.56 | 59.09 |
| ReviewRAG / Qwen-3-8B | 82.87 | 68.26 | 55.81 | 82.52 | 75.92 | 62.96 |
| AspectRAG / GPT-4o-mini | 90.89 | **72.40** | **62.53** | 87.48 | 79.90 | **69.83** |
| AspectRAG / Qwen-3-4B | 87.66 | 67.36 | 56.52 | 87.23 | 75.30 | 63.13 |
| AspectRAG / Qwen-3-8B | **93.47** | 69.96 | 59.27 | **88.38** | **80.68** | 68.33 |

evidente. Esse comportamento é esperado, pois termos de aspecto normalmente correspondem a entidades ou atributos mais explícitos no texto, enquanto termos de opinião tendem a apresentar maior variabilidade linguística e dependência de contexto, tornando sua identificação mais desafiadora. Por exemplo, no ReLi, o Qwen-3-4B em Zero Shot obtém 42.20% em ATE, mas apenas 23.30% em ASTE, indicando que a ausência de exemplos relevantes dificulta a associação entre aspectos e opiniões. Com o Few Shot, ocorre melhora em ATE, mas o desempenho em OTE e ASTE permanece baixo para modelos menores, como o Qwen-3-4B que atinge apenas 13.25% em ASTE. Indicando que apenas fornecer exemplos no prompt não é suficiente para capturar corretamente as relações entre aspectos e opiniões, especialmente em modelos menores.

A introdução da recuperação no ReviewRAG melhora substancialmente esses resultados. No ReHol, por exemplo, o Qwen-3-8B passa de 47.60%

em ASTE no Zero Shot para 62.96% no ReviewRAG. Ainda assim, o uso de trechos longos das avaliações reduz a especificidade das evidências e limita o ganho. Esse comportamento explica por que o ReviewRAG supera o Few Shot, mas continua abaixo do *AspectRAG* em todos os casos.

O *AspectRAG* apresenta os melhores resultados entre todas as variantes. No ReLi, o Qwen-3-8B alcança 59.27% em ASTE, superando tanto o ReviewRAG (55.81%) quanto o Zero Shot (27.19%). No ReHol, a diferença é ainda mais clara: 68.33% em ASTE contra 62.96% do ReviewRAG e 47.60% do Zero Shot. O mesmo padrão aparece no GPT-4o-mini, cuja performance cresce de 55.75% (Zero Shot) para 57.28% (Few Shot), 59.87% (ReviewRAG) e 62.53% (AspectRAG) no ReLi.

Esses resultados mostram que a recuperação orientada a aspectos é o principal fator responsável pelo ganho de qualidade, funcionando como o mecanismo que permite aos modelos gerar triplas mais coerentes e completas mesmo sem treinamento supervisionado.

## 6 Conclusões e trabalhos futuros

Este trabalho apresentou o AspectRAG, uma arquitetura de Recuperação e Geração projetada para as subtarefas de ABSA, com foco na extração de triplas ASTE em português. A proposta integra recuperação semântica guiada por aspectos com um modelo gerativo que utiliza exemplos altamente relevantes para identificar termos de aspecto, termos de opinião e polaridades. Os resultados experimentais mostram que o AspectRAG alcança desempenho superior aos métodos supervisionados avaliados, incluindo OTE-MTL, CMLA-MTL e BOTE, tanto em ATE quanto em OTE e ASTE. Além disso, o modelo se mostrou eficiente mesmo quando implementado com LLMs compactos, indicando que os ganhos obtidos decorrem principalmente da qualidade da recuperação e não apenas da capacidade do modelo gerador.

A variação do parâmetro K revelou que valores intermediários e altos tendem a produzir os melhores resultados, especialmente em domínios com maior diversidade de padrões linguísticos. O estudo de ablação confirmou a importância da etapa de recuperação semântica. Abordagens que não utilizam recuperação, como Zero Shot e Few Shot tradicionais, apresentam desempenho inferior, enquanto o ReviewRAG, embora mais robusto, ainda não atinge o nível obtido pelo AspectRAG. Esses achados indicam que a recuperação orientada por aspectos é um elemento essencial para aumentar a precisão e a consistência das triplas geradas.

Direções futuras incluem o desenvolvimento de estratégias de recuperação mais especializadas, possivelmente com índices diferenciados para tipos variados de evidências, possibilitando recuperação baseada no termo de opinião, por exemplo, além da adoção de técnicas de re-ranqueamento semântico para melhorar a qualidade das evidências recuperadas. Também pretendemos verificar o impacto que o ajuste fino dos modelos para a tarefa de ASTE terá no AspectRAG.

## Referências


José Meléndez Barros e Glauber De Bona. 2021. A deep learning approach for aspect sentiment triplet extraction in portuguese. Em *Intelligent Systems*, páginas 343–358, Cham. Springer International Publishing.

V. Câmara e T. Melo. 2022. Estudo de método de extração de aspectos para português do brasil baseado em regras. *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, páginas 192–203.

Abir Chakraborty. 2024. End-to-end aspect based sentiment analysis using graph attention network. Em *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, página 3663–3668, New York, NY, USA. Association for Computing Machinery.

Z. Chen e T. Qian. 2020. Enhancing aspect term extraction with soft prototypes. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, páginas 2107–2117.

Judith A. Chevalier e Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354.

Keith Cortis e Brian Davis. 2021. Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, 54(7):4873–4965.

Vanessa Câmara e Tiago Melo. 2022. Estudo de método de extração de aspectos para português do brasil baseado em regras. Em *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, páginas 192–203, Porto Alegre, RS, Brasil. SBC.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, e Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. Em *Proceedings of the 37th International Conference on Machine Learning (ICML) / PMLR 119*, páginas 3929–3938. Preprint disponível em arXiv:2002.08909.

Gautier Izacard e Édouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint*,

arXiv:2007.01282. Versão publicada e discussões em conferências posteriores (ex.: EACL/ACL).

S. Y. B. Jian, T. Nayak, N. Majumder, e S. Poria. 2021. Aspect sentiment triplet extraction using reinforcement learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 1–10.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, e Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. Em *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 6769–6781. Preprint disponível em arXiv:2004.04906.

P. Lewis, E. Perez, A. Piktus, F. Petroni, and 1 others. 2020a. Retrieval-augmented generation for knowledge-intensive nlp. Em *Advances in Neural Information Processing Systems (NeurIPS)*, páginas 9459–9474.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, e Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp. Em *Advances in Neural Information Processing Systems (NeurIPS) 33*, páginas 9459–9474. Preprint disponível em arXiv:2005.11401.

X. Li, L. Wang, P. Li, e Z. Yang. 2018. Aspect term extraction with history attention and selective fusion. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.

Bing Liu e Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, páginas 415–463. Springer US, Boston, MA.

M. T. Machado, T. A. S. Pardo, E. E. S. Ruiz, e A. Di Felippo. 2021a. Learning rules for automatic identification of implicit aspects in portuguese. Em *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*.

Mateus Tarcinalli Machado, Thiago Alexandre Salgueiro Pardo, Evandro Eduardo Seron Ruiz, e Ariani Di Felippo. 2021b. Learning rules for automatic identification of implicit aspects in portuguese. Em *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. Disponível também em ResearchGate (PDF).

I. Naglik, L. Zhang, e B. Liu. 2024. Aste-transformer: Modelling dependencies in aspect-sentiment-opinion triplet extraction. *Findings of the Association for Computational Linguistics: EMNLP 2024*, páginas 129–139.

Daniel A. Pereira. 2020. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Gabriel Pereira, Luciano Barbosa, Johny Moreira, Tiago Melo, e Altigran Silva. 2024. Enhancing aspect-based sentiment analysis for portuguese using instruction tuning. Em *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, páginas 990–1001, Porto Alegre, RS, Brasil. SBC.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, and 1 others. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. Em *International workshop on semantic evaluation*, páginas 19–30.

Eloize Seno, Lucas Silva, Fábio Anno, Fabiano Rocha, e Helena Caseli. 2024. Aspect-based sentiment analysis in comments on political debates in Portuguese: evaluating the potential of ChatGPT. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, páginas 312–320, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tun Thura Thet, Jin-Cheon Na, e Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.

Qi Wang e et al. 2024. In-context example retrieval from multi-perspectives for few-shot aspect-based sentiment analysis. Em *Proceedings of LREC-COLING 2024 (LREC Main Conference)*, páginas —. Recuperação de exemplos para prompting / ICL no contexto de ABSA.

W. Wang, S. J. Pan, D. Dahlmeier, e X. Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. Em *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kyung Hyan Yoo e Ulrike Gretzel. 2008. What motivates consumers to write online travel reviews? *Information Technology & Tourism*, 10(4):283–295.

Guoxin Yu, Redmond Liu, Haiyun Jiang, Shuming Shi, e et al. 2023. Retrieval-based aspect sentiment triplet extraction via label interpolation. Em *Findings of the Association for Computational Linguistics: ACL 2023*, páginas —. Código/ PDF: https://aclanthology.org/2023.findings-acl.303.pdf.

C. Zhang, Q. Li, D. Song, e B. Wang. 2020. A multi-task learning framework for opinion triplet extraction. Em *Findings of the Association for Computational Linguistics: EMNLP 2020*, páginas 819–828.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, e Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.