Agrupamentos de Vídeos Relacionados ao Desenvolvimento de Software por Meio de LLMs

Isadora Pereira Marques Martins

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2026

Isadora Pereira Marques Martins

Agrupamentos de Vídeos Relacionados ao Desenvolvimento de Software por Meio de LLMs

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Sistemas de Informação

Orientador: Prof. Dr. Adriano Mendonça Rocha

Monte Carmelo - MG
2026

Dedico este trabalho à minha família, pelo apoio constante e pela presença fundamental ao longo da minha trajetória acadêmica. Dedico-o também a mim mesma, pela perseverança diante dos desafios enfrentados durante a graduação, bem como aos amigos que estiveram ao meu lado ao longo desse processo.

Agradecimentos

Agradeço primeiramente a Deus, por me conceder saúde, força e perseverança para superar os desafios enfrentados ao longo da graduação e alcançar esta importante etapa da minha formação acadêmica.

Agradeço à minha família, pelo apoio incondicional, incentivo constante e compreensão durante todo esse percurso. Em especial, agradeço à minha irmã Rhadyja, que sempre me incentivou e acreditou em mim, mesmo nos momentos mais difíceis.

Agradeço ao meu orientador, Adriano Mendonça Rocha, pela orientação, disponibilidade, paciência e pelas contribuições ao longo do desenvolvimento deste trabalho, que foram essenciais para a sua realização.

Aos professores que fizeram parte da minha trajetória acadêmica, expresso minha sincera gratidão pelos ensinamentos, pela dedicação e pelo conhecimento compartilhado ao longo da graduação, contribuindo de forma significativa para minha formação pessoal e profissional.

"Eu plantei, Apolo regou; mas Deus deu o crescimento. De modo que nem o que planta é alguma coisa, nem o que rega, mas Deus, que dá o crescimento."
(1 Coríntios 3:6–7)

Resumo

O crescimento acelerado do volume de vídeos educacionais em plataformas digitais, como o *YouTube*, torna necessária a adoção de métodos automáticos capazes de organizar e agrupar conteúdos de forma eficiente e semanticamente consistente. Nesse contexto, este trabalho tem como objetivo comparar a eficácia de abordagens baseadas em grandes modelos de linguagem (*Large Language Models* (LLMs)) com métodos tradicionais de clusterização baseados em modelagem de tópicos no agrupamento semântico de transcrições de vídeos relacionados ao desenvolvimento de *software*. A metodologia adotada possui caráter exploratório e comparativo, utilizando transcrições textuais de vídeos educacionais como base de análise. Foram avaliadas duas abordagens principais: métodos tradicionais de clusterização baseados em modelagem de tópicos com *Latent Dirichlet Allocation* (LDA) e abordagens baseadas em LLMs, incluindo diferentes modelos e configurações. Os agrupamentos gerados foram comparados por meio de métricas externas de avaliação de clusterização, especificamente *Adjusted Rand Index* (ARI), *Adjusted Mutual Information* (AMI), *V-Measure* e índice de Jaccard, utilizando como referência adicional uma clusterização manual realizada por um avaliador humano. Os resultados indicam que modelos baseados em LLMs, especialmente aqueles com maior capacidade contextual, produzem agrupamentos semanticamente mais consistentes e com maior alinhamento em relação à referência humana quando comparados a métodos tradicionais utilizados com configurações padrão. Observou-se também que abordagens clássicas, como o LDA, apresentam desempenho competitivo quando adequadamente parametrizadas, embora sejam mais sensíveis ao pré-processamento e à escolha de parâmetros. Conclui-se que a utilização de grandes modelos de linguagem constitui uma alternativa viável e promissora para tarefas de agrupamento semântico de transcrições textuais, oferecendo maior robustez na captura de relações temáticas complexas. Os resultados reforçam o potencial dessas abordagens como ferramentas de apoio à organização automática de conteúdos educacionais e indicam caminhos para pesquisas futuras na área.

Palavras-chave: Clusterização, Grandes Modelos de Linguagem, Processamento de Linguagem Natural, Transcrições de Vídeos, Análise Semântica.

Lista de ilustrações

Lista de tabelas

Lista de siglas

IA Inteligência Artificial

LLMs *Large Language Models* - Linguagens de Grande Porte

LDA *Latent Dirichlet Allocation* - Alocação Latente de Dirichlet

LDADE *Latent Dirichlet Allocation with Differential Evolution* - Alocação Latente de Dirichlet com Evolução Diferencial

DE *Differential Evolution* - Evolução Diferencial

PLN Processamento de Linguagem Natural

GPT *Generative Pre-trained Transformer* - Transformador Generativo Pré-treinado

RNNs *Recurrent Neural Network* - Redes Neurais Recorrentes

LSTMs *Long Short-Term Memory* - Memória de Curto e Longo Prazo

BoW *Bag-of-Words* - Sacola de Palavras

TF-IDF *Term Frequency-Inverse Document Frequency* - Frequência do Termo-Frequência Inversa do Documento

LSH *Locality-Sensitive Hashing*

CNNs *Convolutional Neural Network* - Redes Neurais Convolucionais

PIC *Power Iteration Clustering*

ARI *Adjusted Rand Index*

RI *Rand Index*

AMI *Adjusted Mutual Information*

MI *Mutual Information*

MoE *Mixture-of-Experts* - Mistura de Especialistas

MTP *Multi-Token Prediction*

Sumário

Capítulo 1

Introdução

A utilização da Inteligência Artificial (IA) no desenvolvimento de *software* tem transformado significativamente a forma como grandes volumes de dados são processados e analisados. Com a crescente produção de dados digitais, surge a necessidade de métodos eficientes para organizar e categorizar informações de maneira automática, especialmente em ambientes que concentram grande volume de conteúdo multimídia, como plataformas de compartilhamento de vídeos.

Nesse contexto, o agrupamento automático de conteúdos tem ganhado destaque por permitir a identificação de padrões e similaridades entre diferentes itens, facilitando tarefas como organização, busca e recomendação de informações. Em plataformas como o *YouTube*, embora os vídeos sejam o elemento central, uma parcela relevante do conteúdo pode ser representada de forma textual por meio de metadados e, principalmente, transcrições de áudio. Essas transcrições constituem uma fonte rica de informação semântica e viabilizam a aplicação de técnicas consolidadas de Processamento de Linguagem Natural (PLN) para análise e agrupamento temático.

Os grandes modelos de linguagem, denominados LLMs, baseiam-se em arquiteturas de redes neurais profundas treinadas sobre grandes volumes de dados textuais, o que lhes confere a capacidade de capturar relações semânticas complexas e dependências contextuais de longo alcance. Essas características tornam os LLMs particularmente adequados para tarefas que exigem compreensão semântica refinada, como sumarização, classificação e agrupamento de textos extensos. Assim, no contexto da clusterização de vídeos, os LLMs podem ser empregados de forma indireta, atuando sobre as transcrições textuais dos conteúdos audiovisuais, permitindo uma análise semântica mais profunda do material disponível.

Diferentemente de abordagens que exploram múltiplas modalidades de dados, este trabalho concentra-se exclusivamente na análise textual derivada de vídeos, em especial nas transcrições disponibilizadas pela própria plataforma. Essa escolha permite isolar o impacto das técnicas de processamento textual e avaliar de forma mais controlada o potencial dos LLMs em comparação com métodos tradicionais de clusterização baseados

em modelagem de tópicos e fundamentados em estatísticas de frequência de termos.

1.1 Motivação

A rápida evolução das tecnologias digitais e a crescente disponibilidade de dados multimídia têm transformado a forma como as informações são produzidas, consumidas e analisadas. Com bilhões de horas de vídeos disponibilizados diariamente, plataformas como o *YouTube* tornaram-se não apenas grandes repositórios de conteúdo, mas também ambientes nos quais a organização e a categorização automáticas são essenciais para garantir acesso eficiente à informação relevante.

Nesse cenário, algoritmos de recomendação exercem papel central ao mediar a interação entre usuários e conteúdos, organizando vídeos com base em padrões de comportamento, associações e vínculos estabelecidos a partir das ações dos usuários (GILLESPIE, 2014). Em particular, o algoritmo do *YouTube* adapta continuamente a apresentação dos conteúdos às preferências individuais, facilitando a descoberta de novos vídeos e influenciando a forma como o conhecimento é consumido (SANCHES, 2023).

Contudo, os mecanismos tradicionais de organização e agrupamento de conteúdos textuais derivados de vídeos, como transcrições, nem sempre são capazes de capturar adequadamente relações semânticas mais profundas. Nesse contexto, a utilização de IA e, em especial, de grandes modelos de linguagem, surge como uma alternativa promissora para aprimorar o agrupamento temático automático, oferecendo representações textuais mais ricas e semanticamente estruturadas.

1.2 Problema

O grande volume de vídeos disponíveis em plataformas como o *YouTube* gera um vasto conjunto de dados textuais provenientes de transcrições, cuja organização temática automática ainda representa um desafio. Métodos tradicionais de clusterização baseados em modelagem de tópicos e fundamentados em estatísticas de frequência de termos, podem apresentar limitações na captura de relações semânticas complexas presentes em textos longos e contextualizados.

Diante desse cenário, surge a seguinte questão de pesquisa: *em que medida os grandes modelos de linguagem podem aprimorar o agrupamento semântico de transcrições de vídeos educacionais?*

1.3 Hipótese

Parte-se da hipótese de que abordagens baseadas em grandes modelos de linguagem (LLMs) produzem agrupamentos semanticamente mais consistentes de transcrições de ví-

deos quando comparadas a métodos tradicionais de clusterização baseados em modelagem de tópicos. Essa hipótese fundamenta-se na capacidade dos LLMs de gerar representações textuais contextualizadas, permitindo uma melhor captura de relações temáticas complexas e de estruturas semânticas subjacentes ao conteúdo.

1.4 Objetivos

O objetivo geral desta pesquisa é comparar a eficácia de abordagens baseadas em grandes modelos de linguagem (LLMs) com métodos tradicionais de clusterização baseados em modelagem de tópicos no agrupamento semântico de transcrições de vídeos educacionais, utilizando métricas externas de avaliação de clusterização e comparação com uma referência humana. A fim de alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- ❏ Comparar o desempenho de diferentes LLMs na tarefa de clusterização de transcrições de vídeos;
- ❏ Comparar a eficácia das abordagens baseadas em LLMs com métodos tradicionais, como LDA e *Latent Dirichlet Allocation with Differential Evolution* (LDADE);
- ❏ Avaliar os agrupamentos gerados por meio de métricas externas de clusterização;
- ❏ Analisar o impacto de duas abordagens distintas utilizando LLMs no processo de clusterização;
- ❏ Comparar o desempenho do LDA com sua variação otimizada, o LDADE.

1.5 Contribuições

Este trabalho contribui com uma análise comparativa entre métodos tradicionais de clusterização baseados em modelagem de tópicos e abordagens baseadas em LLMs aplicadas ao agrupamento semântico de transcrições de vídeos. A pesquisa apresenta uma avaliação sistemática da qualidade dos agrupamentos gerados por diferentes técnicas, utilizando métricas externas e uma referência humana como aproximação da interpretação semântica, oferecendo evidências quantitativas sobre o desempenho relativo dessas abordagens em tarefas de organização temática automática.

1.6 Organização da Monografia

Esta monografia está organizada da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica e os trabalhos relacionados; o Capítulo 3 descreve os experimentos

realizados e a análise dos resultados obtidos; e o Capítulo 4 apresenta as conclusões do trabalho e as possibilidades de pesquisas futuras.

Capítulo 2

Fundamentação Teórica

Neste capítulo, será apresentada a base teórica que sustenta o desenvolvimento deste trabalho, abordando os conceitos fundamentais para o entendimento da abordagem proposta.

2.1 PLN

O PLN é uma subárea da inteligência artificial (IA) e da linguística computacional, focada na interação entre computadores e humanos por meio da linguagem natural. O objetivo do PLN é possibilitar que máquinas entendam, interpretem, e gerem texto de maneira que seja natural para humanos. Em uma decomposição mais detalhada, o PLN pode ser subdividido em cinco estágios principais de análise (OLIVEIRA et al., 2020):

- Tokenização: tem por objetivo dividir o texto em unidades menores, chamados de *tokens*, podendo ser dividido em:
  - Tokenização por palavras: divide o texto em palavras individuais, por exemplo: "A tokenização é importante!", os *tokens* seriam: ["A", "tokenização", "é", "importante", "!"];
  - Tokenização por frases: divide o texto em frases completas, por exemplo: "A tokenização é importante! Ela facilita o processamento.", os *tokens* seriam: ["A tokenização é importante!", " Ela facilita o processamento."];
  - Tokenização por caracteres: trata de cada caractere como um único *token*, por exemplo: "texto", os *tokens* seriam: ["t", "e", "x", "t", "o"].
- Análise léxica: relaciona os *tokens* às suas categorias léxicas e formas básicas, como identificar variantes morfológicas (como verbos conjugados) e associá-las aos seus lemas (formas base). Por exemplo, no processo de tokenização de uma frase como "Os gatos correm rapidamente", os *tokens* são dados pela tokenização por caracteres. A análise léxica categorizaria "gatos" como um substantivo plural, "correm" como

uma forma conjugada do verbo “correr”, e associaria cada palavra ao seu lema (no caso, “gato”, “correr”).

- ❏ Análise sintática: concentra-se no relacionamento entre as palavras, atribuindo a cada uma seu papel estrutural dentro das frases, além de analisar como as frases podem se combinar para formar sentenças mais complexas. Por exemplo, a frase “O menino viu o cachorro no parque”. A análise sintática identificaria “O menino” como o sujeito, “viu” como o verbo (predicado) e “o cachorro no parque” como o objeto direto. Além disso, a análise verificaria a estrutura da frase e como as palavras se relacionam para formar uma sentença gramaticalmente correta.

- ❏ Análise semântica: foca no significado das palavras, expressões fixas e sentenças completas. Por exemplo, em uma frase como “O cachorro está latindo”, a análise semântica identificaria que “cachorro” refere-se a um animal doméstico, e “latindo” é uma ação associada ao som que o cachorro faz. A análise se concentraria em entender o significado das palavras e como elas contribuem para o significado da sentença como um todo.

- ❏ Análise pragmática: busca interpretar uma frase no contexto, levando em conta referências pronominais e a coerência entre as frases adjacentes. Por exemplo, considere a frase “Maria não veio para o jantar. Ela estava doente.”, quando tokenizada por frases, a análise pragmática verificaria a referência pronominal “Ela” e associaria essa palavra a “Maria” na frase anterior. Além disso, essa análise consideraria o contexto para garantir que a segunda frase seja coerente com a anterior, inferindo que Maria estava doente, razão pela qual não compareceu ao jantar.

2.2 Arquitetura *Transformer*

Os *Transformers* são uma arquitetura de redes neurais introduzida por Vaswani et al. (2017), que revolucionou o campo do PLN. A principal inovação do modelo é o mecanismo de *self-attention*, que permite ao modelo atribuir pesos a diferentes palavras em uma sentença, independentemente de sua posição na sequência, contrastando com arquiteturas tradicionais baseadas em recorrência.

Com o avanço das aplicações de Processamento de Linguagem Natural (PLN) e o aumento do volume de dados disponíveis, tornou-se necessária a adoção de arquiteturas capazes de modelar dependências de longo alcance de forma eficiente e escalável. Nesse contexto, modelos baseados em redes recorrentes, como *Recurrent Neural Network* (RNNs) e *Long Short-Term Memory* (LSTMs), passaram a apresentar limitações relacionadas ao processamento sequencial e à dificuldade em capturar relações distantes no texto. A arquitetura *Transformer* foi proposta justamente para superar essas limitações, utilizando

mecanismos de atenção como componente central e permitindo maior paralelismo durante o treinamento.

A Figura 1 enfatiza a arquitetura do *Transformer decoder-only*, cujo objetivo é prever o próximo *token* ou palavra de uma sequência a partir de um *prompt* fornecido como entrada. Esse modelo analisa os *tokens* anteriores da sequência para gerar a continuação mais provável, produzindo frases coerentes com base nos padrões aprendidos durante o treinamento sobre grandes volumes de dados (BOUCHARD, 2024). Alguns de seus pontos principais são:

- ❏ Tokenização e *Embeddings*: transformam o texto de entrada em representações vetoriais contínuas, permitindo que o modelo processe semanticamente as palavras da sequência.
- ❏ *Positional Encoding*: adiciona informações sobre a posição dos *tokens* na sequência, permitindo ao modelo capturar a ordem das palavras, uma vez que o mecanismo de atenção não possui noção posicional inerente.
- ❏ *Multi-Head Self-Attention*: permite ao modelo focar em várias partes da entrada simultaneamente, capturando relações contextuais entre as palavras. Este é o elemento-chave que possibilita a modelagem de dependências longas no texto.
- ❏ Redes *Feed-Forward*: são redes totalmente conectadas que processam as representações das palavras individualmente, aplicando transformações não lineares para gerar representações mais refinadas.
- ❏ Conexões Residuais e Normalização: auxiliam na estabilização do treinamento e no fluxo de gradientes, permitindo o empilhamento de múltiplas camadas do modelo.
- ❏ Camada Linear e Softmax: projetam a saída do modelo para o espaço do vocabulário e produzem uma distribuição de probabilidade sobre os possíveis próximos *tokens*.

A Figura 2 destaca a arquitetura do *Transformer encoder-decoder*, que é composta por múltiplas camadas de codificadores e decodificadores, sendo que cada camada possui dois mecanismos principais:

- ❏ Separação entre codificador e decodificador: o codificador é responsável por processar toda a sequência de entrada e gerar representações contextuais, enquanto o decodificador utiliza essas representações para gerar a sequência de saída.
- ❏ Atenção cruzada (*Encoder–Decoder Attention*): o decodificador possui um mecanismo de atenção que permite consultar diretamente as representações produzidas pelo codificador, alinhando a entrada com a saída gerada.

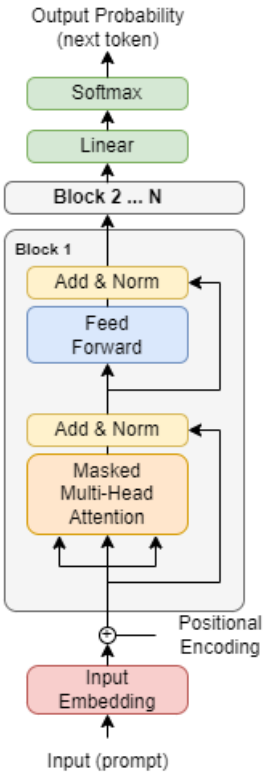

Figura 1 – Arquitetura *Transformer decoder-only*. Fonte: Bouchard (2024)

- ❏ Atenção mascarada no decodificador: o uso de *Masked Multi-Head Attention* impede que o modelo utilize informações futuras durante a geração da sequência.
- ❏ Processamento paralelo no codificador: todas as palavras da sequência de entrada são processadas simultaneamente, aumentando a eficiência computacional e a qualidade das representações contextuais.
- ❏ Adequação para tarefas de transformação de sequência: essa arquitetura é amplamente utilizada em tarefas como tradução automática, sumarização e outras aplicações de geração condicionada de texto.

Destaca-se que o codificador processa a sequência de entrada e gera representações intermediárias ricas em informações contextuais, enquanto o decodificador utiliza essas representações para gerar a sequência de saída, seja na forma de tradução, sumarização ou outras tarefas de linguagem.

Além disso, o mecanismo de atenção mascarada (*Masked Multi-Head Attention*) no decodificador impede que o modelo "veja" as palavras futuras durante a geração de texto, garantindo que ele utilize apenas palavras passadas.

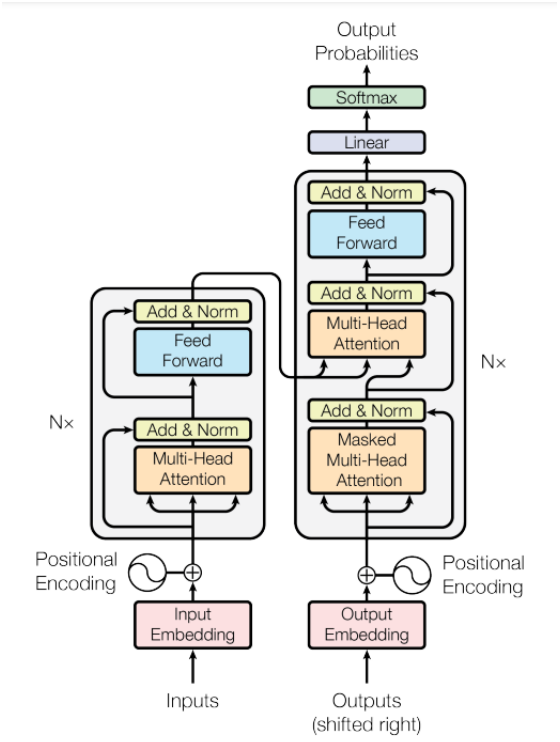

Figura 2 – Arquitetura *Transformer encoder-decoder*. Fonte: Vaswani et al. (2017)

Por sua natureza paralelizável, os *Transformers* têm se mostrado altamente eficientes para treinamento em grandes volumes de dados, o que levou ao desenvolvimento de grandes modelos de linguagem, como o GPT que utiliza o modelo *Transformer decoder-only*. Acrescenta-se que esses modelos têm apresentado excelente desempenho em uma variedade de tarefas de PLN.

2.3 Arquitetura MoE

O modelo *Mixture-of-Experts* (MoE) consiste na utilização de múltiplos *experts* especializados, cada um responsável por capturar diferentes padrões ou subespaços dos dados. Essa abordagem é baseada na arquitetura *Transformer*, diferenciando-se por incorporar um mecanismo adicional de roteamento, responsável por selecionar dinamicamente quais *experts* serão ativados para o processamento de cada *token*.

No MoE, cada especialista é tipicamente implementado como uma rede *Feed-Forward* independente. Para determinar quais *tokens* serão encaminhados a cada especialista, utiliza-se uma rede de *gate* (ou roteador), que calcula uma distribuição de probabilidade sobre os *experts* disponíveis. Apenas um subconjunto desses especialistas é ativado para cada *token*, caracterizando um mecanismo de ativação esparsa.

Diferentemente da arquitetura *Transformer* tradicional, na qual a camada *Feed-Forward*

é densa e aplicada integralmente a todos os *tokens*, o MoE substitui essa etapa por múltiplos especialistas paralelos. Conforme descrito por Fedus, Zoph e Shazeer (2021), essa estratégia permite aumentar significativamente o número total de parâmetros do modelo, mantendo aproximadamente constante o custo computacional por *token*, uma vez que apenas um número reduzido de especialistas é efetivamente ativado durante a inferência.

Segundo Fedus, Zoph e Shazeer (2021), o termo *expert* refere-se a um componente especializado em capturar padrões específicos presentes nos dados, de modo complementar aos demais especialistas. Esse mecanismo possibilita maior capacidade representacional sem exigir a ativação simultânea de todos os parâmetros do modelo.

A Figura 3 ilustra a arquitetura *Mixture-of-Experts* (MoE), uma extensão da arquitetura *Transformer* que abrange os principais componentes:

- Especialistas (*Experts*): consistem em múltiplas redes neurais independentes, geralmente implementadas como redes *Feed-Forward*, responsáveis por capturar diferentes padrões ou subestruturas presentes nos dados.
- Rede de *Gate* (Roteador): mecanismo responsável por decidir, para cada *token*, quais especialistas serão ativados, permitindo a seleção dinâmica de um subconjunto de especialistas.
- Ativação esparsa: ao contrário das camadas densas tradicionais, apenas alguns especialistas são ativados para cada *token*, reduzindo o custo computacional e possibilitando a escalabilidade do modelo.
- Combinação das saídas dos especialistas: as saídas dos especialistas selecionados são combinadas, geralmente por meio de uma soma ponderada, produzindo a representação final do *token*.
- Integração com o *Transformer*: o mecanismo MoE substitui a camada *Feed-Forward* tradicional do *Transformer*, mantendo inalterados os mecanismos de atenção, bem como as conexões residuais e a normalização.

Adicionalmente, arquiteturas MoE geralmente incorporam mecanismos de balanceamento de carga, cujo objetivo é evitar que apenas um pequeno subconjunto de especialistas seja excessivamente ativado durante o processamento (SHAZEER et al., 2017). Esses mecanismos incentivam uma distribuição mais uniforme dos *tokens* entre os especialistas, contribuindo para um uso mais eficiente da capacidade do modelo e para a estabilidade do treinamento.

Durante o treinamento, o mecanismo de roteamento aprende a associar padrões específicos de entrada aos especialistas mais adequados, enquanto, na fase de inferência, esse roteamento possibilita manter alta capacidade representacional com custo computacional reduzido. Apesar dessas vantagens, arquiteturas MoE também introduzem desafios

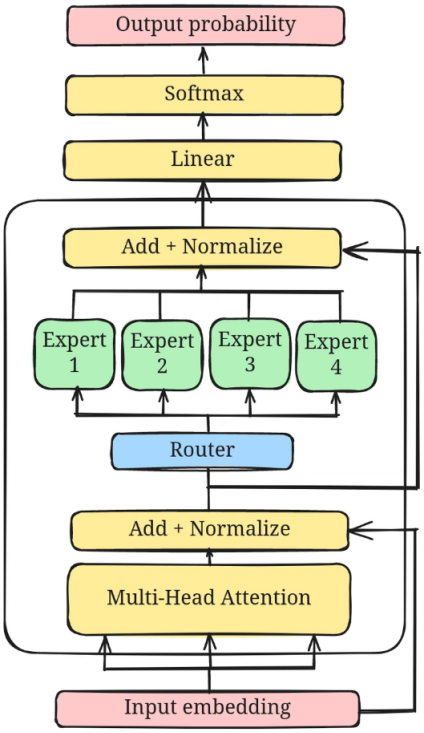

Figura 3 – Arquitetura MoE. Fonte: Oriol (2024)

adicionais, como maior complexidade de implementação e a necessidade de mecanismos auxiliares para controle do roteamento e do balanceamento entre especialistas (DU et al., 2022).

2.4 Mecanismo MTP

O mecanismo de predição especulativa de múltiplos *tokens* consiste em uma estratégia de geração textual na qual múltiplos *tokens* candidatos podem ser propostos em um único ciclo de inferência, em contraste com a geração estritamente sequencial de um *token* por vez. Essa abordagem busca reduzir a latência da geração ao permitir que mais de uma posição da sequência seja validada simultaneamente, mantendo a natureza autoregressiva do modelo principal.

Diferentemente da predição tradicional, na qual cada *token* é confirmado individualmente antes da geração do próximo, a estratégia descrita por Leviathan, Kalman e Matias (2022) utiliza um modelo auxiliar para propor múltiplos *tokens*, que são posteriormente verificados pelo modelo maior. Caso as predições sejam aceitas, múltiplos *tokens* são incorporados à sequência em um único passo, reduzindo o número de chamadas ao modelo principal e, consequentemente, o tempo de inferência.

Segundo Leviathan, Kalman e Matias (2022), essa técnica pode proporcionar ganhos

significativos de eficiência computacional e redução de latência, preservando a qualidade da geração textual. Contudo, os benefícios dependem da compatibilidade entre o modelo auxiliar e o modelo principal, bem como do domínio de aplicação.

Entretanto, essa abordagem também apresenta limitações. Os ganhos de desempenho dependem diretamente da qualidade do modelo auxiliar utilizado para gerar os *tokens* candidatos. Caso o modelo auxiliar produza predições pouco alinhadas ao modelo principal, a taxa de rejeição pode aumentar, reduzindo os benefícios em termos de latência. Além disso, a eficiência obtida varia conforme o tamanho relativo dos modelos envolvidos e o domínio de aplicação (LEVIATHAN; KALMAN; MATIAS, 2022).

A Tabela 1 apresenta uma comparação entre a predição de *token* único e a predição de múltiplos *tokens*, com base nas descrições apresentadas por Liora (2025).

Tabela 1 – Diferença entre Predição de *Token* Único e Predição de Múltiplos *Tokens* (adaptado de Liora (2025))

Critério	**Predição de Token Único**	**Predição de Múltiplos Tokens**
Modo de Geração	Geração de um token por vez, baseada nos tokens anteriores	Geração de múltiplos tokens simultaneamente em uma única etapa
Exemplos de Modelos	GPT-2 e modelos autoregressivos clássicos	Modelos recentes de grande porte (por exemplo, GPT-4, Claude, Gemini)
Velocidade de Processamento	Mais lenta, devido à dependência sequencial entre tokens	Potencialmente mais rápida, devido à geração paralela de tokens
Coerência Geral	Pode apresentar degradação em textos longos, com maior risco de repetições	Pode favorecer maior coerência semântica e gramatical
Antecipação de Contexto	Limitada, com menor visão global da sequência	Maior consideração do contexto global
Fluência da Geração	Pode resultar em construções mais artificiais ou truncadas	Tende a produzir geração mais natural e fluida

Ressalta-se que, no caso de modelos proprietários, detalhes completos sobre a implementação interna das estratégias de predição nem sempre são publicamente divulgados, o que limita afirmações categóricas sobre o uso específico do MTP. A partir da Tabela 1, observa-se que a predição de múltiplos *tokens* constitui uma alternativa à geração estritamente autoregressiva, oferecendo ganhos potenciais em eficiência e fluidez textual. Contudo, tais vantagens dependem da forma como a estratégia é integrada ao modelo e do contexto de aplicação, não configurando uma substituição universal à predição de *token* único.

A Tabela 2 sintetiza a relação hierárquica e funcional entre a arquitetura *Transformer*, a extensão MoE e a estratégia MTP. O *Transformer* constitui a arquitetura base do

modelo, definindo sua estrutura fundamental e os mecanismos de atenção responsáveis pela modelagem de dependências de longo alcance.

O MoE, por sua vez, atua como uma extensão arquitetural do *Transformer*, modificando especificamente a camada *Feed-Forward* ao introduzir múltiplos especialistas e um mecanismo de roteamento, com o objetivo de aumentar a capacidade do modelo sem elevar proporcionalmente o custo computacional.

Já o MTP não altera a estrutura arquitetural do modelo, atuando no processo de predição ao permitir a geração simultânea de múltiplos *tokens*. Dessa forma, o MTP pode ser aplicado tanto a modelos *Transformer* tradicionais quanto àqueles que incorporam MoE, contribuindo principalmente para a redução da latência e para ganhos de eficiência durante a inferência.

Tabela 2 – Relação conceitual entre Transformer, MoE e MTP

Aspecto	**Transformer**	**MoE**	**MTP**
Tipo	Arquitetura base	Extensão arquitetural	Mecanismo/Estratégia
Atuação	Estrutura do modelo	Camada *Feed-Forward*	Processo de predição
Modifica a arquitetura	Sim	Sim	Não
Usa atenção	Sim	Sim	Sim (indiretamente)
Objetivo principal	Modelar dependências	Aumentar capacidade	Reduzir latência
Pode ser combinado	—	Sim (com Transformer)	Sim (com ambos)

2.5 LLMs

Os LLMs são modelos de aprendizado profundo projetados para compreender, gerar e manipular linguagem natural. O termo “large” (grande) refere-se ao elevado número de parâmetros desses modelos, os quais são ajustados durante o processo de treinamento com o objetivo de capturar padrões e estruturas linguísticas a partir de extensos conjuntos de dados textuais (GOMES, 2024).

Devido ao grande volume de dados utilizado no treinamento, esses modelos são capazes de produzir textos coerentes e fluentes, resultado direto da arquitetura de redes neurais empregada, como os *Transformers*, introduzidos por Vaswani et al. (2017), que revolucionaram o campo do PLN. Além disso, as LLMs apresentam desempenho elevado em diversas tarefas de processamento de linguagem natural, como sumarização, geração e tradução de textos. Modelos como o GPT e o Gemini destacam-se como exemplos representativos dessa classe.

A Tabela 3 apresenta uma visão comparativa entre alguns dos principais modelos de linguagem de grande porte e as arquiteturas discutidas anteriormente neste capítulo, servindo como uma síntese conceitual antes da análise individual de cada modelo.

Tabela 3 – Relação entre LLMs e arquiteturas/mecanismos utilizados

Modelo	**Transformer**	**MoE**	**MTP**
GPT	Sim	Não divulgado	Não divulgado
Gemini	Sim	Parcial	Não divulgado
DeepSeek	Sim	Sim	Sim

2.5.1 ChatGPT

Desenvolvido pela *OpenAI*, o ChatGPT utiliza a arquitetura *Transformer* para interagir de maneira conversacional com os usuários. Projetado para compreender e gerar texto em linguagem natural, o ChatGPT tem sido amplamente empregado em diversas aplicações, desde assistentes virtuais até ferramentas de apoio à criação de conteúdo.

Ao longo de suas diferentes versões, como o GPT-2, GPT-3, GPT-4 e modelos mais recentes da família GPT, cada iteração apresentou avanços significativos em termos de compreensão contextual, fluidez na geração de texto e capacidade de lidar com uma gama mais ampla de tarefas. Esses avanços incluem melhorias na interpretação de perguntas complexas, maior controle sobre o tom e o estilo das respostas, bem como desempenho aprimorado em tarefas de raciocínio e resolução de problemas, conforme relatado pela OpenAI em suas comunicações técnicas (OpenAI, 2023).

O modelo é treinado a partir de grandes volumes de dados textuais, o que lhe permite identificar padrões linguísticos, relações contextuais e nuances semânticas. Essa capacidade de gerar respostas coerentes e relevantes em tempo real decorre do treinamento em extensos conjuntos de dados, que incluem diálogos, textos informativos e outros formatos textuais (OpenAI, 2023).

Além da geração de texto, o ChatGPT é capaz de realizar tarefas como responder a perguntas, fornecer explicações, auxiliar em atividades educacionais e participar de interações conversacionais informais. Essa versatilidade o torna uma ferramenta relevante no campo do processamento de linguagem natural, com aplicações em áreas como atendimento ao cliente, educação, marketing e suporte à tomada de decisão.

Ressalta-se que, embora o ChatGPT seja fundamentado na arquitetura *Transformer*, detalhes completos sobre possíveis extensões arquiteturais, estratégias de treinamento ou métodos específicos de predição empregados nas versões mais recentes do modelo não são integralmente divulgados, em razão do caráter proprietário da tecnologia (OpenAI, 2023).

2.5.2 Gemini

Desenvolvido pelo *Google DeepMind*, o Gemini é um modelo de linguagem de grande porte projetado para compreender, gerar e manipular linguagem natural de forma eficiente. Assim como o ChatGPT, o Gemini é fundamentado na arquitetura *Transformer*, originalmente proposta por Vaswani et al. (2017), o que lhe permite capturar padrões complexos e dependências de longo alcance na linguagem.

Treinado a partir de grandes volumes de dados textuais, o Gemini é capaz de desempenhar diversas tarefas no campo do processamento de linguagem natural, incluindo geração de texto, tradução automática e sumarização de conteúdos. Sua habilidade em compreender contextos amplos e manter diálogos coesos o torna particularmente adequado para aplicações em assistentes virtuais e sistemas conversacionais (Google DeepMind, 2023).

As versões mais recentes do modelo, introduziram avanços significativos em termos de capacidade de processamento e compreensão contextual. Essas versões suportam sequências de entrada substancialmente maiores, possibilitando o tratamento de documentos extensos e contextos complexos sem perda significativa de coerência. Além disso, o Gemini apresentou melhorias em tarefas de raciocínio lógico e inferência, favorecendo interações mais naturais e precisas com os usuários (Google DeepMind, 2023).

De modo geral, o modelo destaca-se por sua capacidade de adaptação a diferentes estilos linguísticos e contextos de uso, oferecendo respostas relevantes e bem estruturadas. Esse desempenho decorre de um treinamento extensivo em dados diversificados, que permite ao Gemini lidar com uma ampla variedade de tópicos e formatos textuais. Ressalta-se, entretanto, que detalhes completos sobre sua arquitetura interna e eventuais extensões ou estratégias específicas de predição não são integralmente divulgados, em função do caráter proprietário do modelo (Google DeepMind, 2023).

2.5.3 DeepSeek

Desenvolvido pela *DeepSeek-AI*, o DeepSeek é um modelo de linguagem de grande porte projetado para compreender, gerar e manipular linguagem natural. Assim como modelos consolidados, como o ChatGPT e o Gemini, o DeepSeek é fundamentado na arquitetura *Transformer*, o que lhe permite capturar padrões complexos e dependências de longo alcance em sequências textuais.

Uma característica explicitamente documentada do DeepSeek é a adoção da abordagem conhecida como *Mixture of Experts* (MoE). Diferentemente de outros modelos de linguagem de grande porte, cujos detalhes arquiteturais e estratégias específicas de predição, como o uso de *Multi-Token Prediction* (MTP) ou mecanismos de especialização, não são integralmente divulgados, o DeepSeek descreve de forma aberta a utilização de múltiplos especialistas especializados em diferentes aspectos do processamento linguístico (DeepSeek-AI, 2024; DeepSeek-AI, 2025).

Nessa arquitetura, um mecanismo de roteamento é responsável por selecionar dinamicamente quais especialistas devem ser ativados para cada entrada, permitindo que apenas uma fração do modelo seja utilizada em cada etapa de inferência. Essa estratégia contribui para maior eficiência computacional e melhor escalabilidade, mantendo desempenho competitivo em tarefas de geração de texto, raciocínio lógico e compreensão contextual (DeepSeek-AI, 2024).

Graças à combinação entre a arquitetura *Transformer* e o mecanismo MoE, o DeepSeek é capaz de realizar tarefas como geração de texto, tradução automática, sumarização e apoio ao raciocínio, produzindo respostas coerentes e contextualmente relevantes. Essa abordagem o torna particularmente interessante em cenários que demandam eficiência no uso de recursos computacionais sem comprometer a qualidade das respostas.

Ressalta-se que, apesar da documentação disponível sobre a adoção da arquitetura MoE, detalhes completos acerca da estrutura interna do modelo, dos critérios de roteamento e das estratégias específicas de treinamento não são integralmente divulgados, em função do caráter proprietário da tecnologia desenvolvida pela *DeepSeek-AI* (DeepSeek-AI, 2025).

2.6 Clusterização

A clusterização é uma técnica de aprendizado não supervisionado que visa agrupar dados semelhantes em conjuntos, ou *clusters*. Essa abordagem é amplamente utilizada em diversas áreas, como mineração de dados, processamento de linguagem natural e análise de imagem, permitindo a identificação de padrões e *insights* que não seriam evidentes em uma análise superficial (JAIN; MURTY; FLYNN, 1999). Por ser uma técnica não supervisionada, a clusterização não requer um conjunto de dados rotulados para ser aplicada; ao invés disso, ela utiliza algoritmos de aprendizado de máquina para identificar padrões subjacentes e relações entre os dados.

De maneira geral, o processo de clusterização envolve a definição de uma medida de similaridade ou distância entre os elementos do conjunto de dados, a partir da qual os algoritmos buscam formar grupos internamente homogêneos e externamente heterogêneos. Diferentes métodos de clusterização adotam estratégias distintas para a formação desses grupos, como a minimização da distância intra-cluster, a maximização da separação entre clusters ou a modelagem probabilística da distribuição dos dados. A escolha do algoritmo e da métrica de similaridade influencia diretamente a qualidade dos agrupamentos obtidos e deve considerar as características do conjunto de dados e os objetivos da análise.

2.6.1 LDA

A LDA é um modelo generativo probabilístico amplamente utilizado para a identificação de tópicos latentes em conjuntos de documentos textuais. Conforme proposto por Blei, Ng e Jordan (2003), o modelo assume que cada documento pode ser representado como uma combinação de múltiplos tópicos, enquanto cada tópico é caracterizado por uma distribuição probabilística sobre o vocabulário. Dessa forma, a LDA busca capturar estruturas temáticas subjacentes em coleções de textos de maneira não supervisionada.

O modelo baseia-se em dois conjuntos principais de variáveis latentes, a distribuição de tópicos associada a cada documento e a distribuição de palavras associada a cada tó-

pico. O processo de inferência da LDA é fundamentado em métodos bayesianos, os quais estimam essas distribuições a partir dos dados observados. Durante o treinamento, o algoritmo analisa o conjunto de documentos e infere quais palavras são mais representativas de cada tópico, permitindo a descoberta automática de padrões temáticos em grandes volumes de texto (GRIFFITHS; STEYVERS, 2004).

No contexto da análise de agrupamentos textuais, a LDA pode ser interpretada como uma forma de *clusterização suave* (*soft clustering*). Diferentemente de métodos de clusterização rígida, nos quais cada documento é atribuído a um único grupo, a LDA associa cada documento a múltiplos tópicos com diferentes probabilidades de pertencimento. Essa representação probabilística permite descrever a composição temática dos documentos de forma mais flexível e informativa, refletindo a natureza multifacetada dos textos.

A principal aplicação da LDA concentra-se em tarefas de modelagem de tópicos, nas quais o modelo auxilia na organização, exploração e interpretação de grandes coleções documentais. Além disso, as distribuições de tópicos geradas pelo modelo podem ser utilizadas como base para etapas posteriores de clusterização ou para a avaliação da similaridade entre documentos, desempenhando um papel relevante em pipelines de análise no contexto do PLN.

2.6.2 LDADE

O LDADE é uma extensão do LDA que incorpora técnicas de otimização baseadas em algoritmos evolutivos para o ajuste automático de seus hiperparâmetros. Diferentemente do LDA tradicional, no qual parâmetros como o número de tópicos e os coeficientes das distribuições de Dirichlet devem ser definidos manualmente, o LDADE emprega o algoritmo de *Differential Evolution* (DE) para explorar um espaço de busca previamente estabelecido, identificando configurações que maximizem métricas de qualidade dos tópicos gerados.

Agrawal, Fu e Menzies (2018) demonstram que o LDA apresenta instabilidade significativa na geração de tópicos, especialmente quando há variações na ordem dos dados de entrada, fenômeno denominado *order effects*. Como solução, os autores propõem o LDADE como uma abordagem de *Search-Based Software Engineering*, utilizando DE para otimizar os parâmetros do modelo e reduzir a instabilidade dos clusters produzidos. Os resultados experimentais indicam melhorias tanto na estabilidade dos tópicos quanto no desempenho em tarefas supervisionadas e não supervisionadas.

Dessa forma, o LDADE representa uma alternativa mais robusta ao LDA tradicional, reduzindo a dependência de ajustes empíricos e aumentando a confiabilidade da modelagem de tópicos em diferentes conjuntos de dados.

Para melhor evidenciar as diferenças estruturais entre o LDA tradicional e o LDADE, a Tabela 4 apresenta uma comparação dos principais hiperparâmetros envolvidos e da estratégia de definição adotada em cada abordagem.

Tabela 4 – Comparação dos hiperparâmetros no LDA e no LDADE

Parâmetro	**Descrição**	**LDA**	**LDADE**
Número de tópicos (K)	Define a quantidade de tópicos latentes a serem extraídos do corpus	Definido manualmente	Otimizado automaticamente via DE
α (documento–tópico)	Controla a distribuição de tópicos por documento	Definido manualmente	Otimizado automaticamente via DE
β (tópico–palavra)	Controla a distribuição de palavras por tópico	Definido manualmente	Otimizado automaticamente via DE
Algoritmo de otimização	Estratégia para ajuste de hiperparâmetros	Não utiliza otimização evolutiva	Utiliza *Differential Evolution*

Observa-se que o modelo probabilístico subjacente permanece o mesmo em ambas as abordagens, sendo a principal diferença o mecanismo de definição dos hiperparâmetros. Enquanto no LDA tradicional esses valores são definidos manualmente, no LDADE são ajustados automaticamente por meio de otimização evolutiva, o que pode impactar diretamente a estabilidade e a qualidade dos tópicos gerados.

Nesse contexto, a utilização do LDA e do LDADE neste trabalho busca estabelecer uma base comparativa sólida frente às abordagens baseadas em LLMs, permitindo analisar diferenças estruturais, estabilidade dos agrupamentos e qualidade semântica dos *clusters* gerados.

2.7 Trabalhos Relacionados

Na literatura, há alguns trabalhos que abordam a clusterização de vídeos do *YouTube*; em sua grande maioria, são trabalhos em inglês. Esta seção visa elencar certos trabalhos que foram usados como orientação para o desenvolvimento deste trabalho. Os artigos buscam identificar padrões ou entender as similaridades dos dados, baseando-se nos dados multimídia como áudio, imagem, vídeos e texto (transcrições).

Basu, Yu e Zimmermann (2016) fazem uso de transcrições de vídeos de palestras com base na similaridade dos tópicos extraídos, utilizando técnicas de modelagem como LDA e *Fuzzy C-Means*. Enquanto o LDA é utilizado para identificar os tópicos principais, o *Fuzzy C-Means* é aplicado para agrupar os vídeos com base nos tópicos. A clusterização *Fuzzy* possibilita que os vídeos façam parte de mais de um *cluster*, o que transparece a ambiguidade e imprecisão dos dados educacionais, melhorando a precisão da categorização. A técnica foi aplicada para agrupar vídeos educativos de plataformas como o *YouTube* e *VideoLectures.net*, com foco em diversas disciplinas acadêmicas.

Basu et al. (2016) abordam um sistema "Videopedia" que integra vídeos educacionais e conteúdo textual de *blogs* e *wikis* para fornecer recomendações de vídeo relevantes que ajudam a explicar os conceitos apresentados nas páginas *web*. O LDA é aplicado para mapear o conteúdo dos vídeos e os textos das páginas da *web* em um espaço semântico comum, que permite que sejam encontrados vídeos relacionados aos tópicos abordados nos textos. O sistema faz a comparação dos tópicos dos vídeos com os tópicos das páginas *web* e, com base na similaridade, vídeos com tópicos semelhantes são recomendados para complementar o aprendizado do usuário.

Jansen et al. (2017) apresentam uma maneira de descobrir eventos de áudios recorrentes e categorizá-los automaticamente, utilizando métodos não supervisionados em um conjunto de 1 milhão de vídeos (aproximadamente 45 mil horas de áudio). Os áudios extraídos são da plataforma do *YouTube* e são rotulados de forma fraca com base em 200 categorias gerais de áudio. O artigo aplica a clusterização não paramétrica (*DenStream*), que se adequa para processamento em tempo real e cria um número dinâmico de *clusters* à medida que os dados evoluem, além de usar *embeddings* neurais de áudio (transformá-los em números para que o computador possa entender) para melhorar a representação das características do áudio. O aprendizado fraco supervisionado associa os *clusters* de áudio aos metadados dos vídeos para localizar eventos de áudio; já a detecção de atividade informativa calcula a relevância semântica dos *clusters* para reduzir o processamento, filtrando eventos que são considerados irrelevantes.

Turcu et al. (2019) apresentam um sistema de busca eficiente para recuperar transcrições de vídeos educacionais, com foco em vídeos em espanhol disponíveis na *Universitat Politècnica de València*. O artigo aborda técnicas de Processamento de Linguagem Natural (PLN), como *Bag-of-Words* (BoW) e *Term Frequency-Inverse Document Frequency* (TF-IDF), técnicas de modelagem de tópicos, como LDA, e técnicas de clusterização, como K-Means e xMeans. Além disso, utiliza a clusterização para agrupar as transcrições em domínios temáticos. Cada transcrição de vídeo é processada, reduzindo as palavras às suas raízes (BoW), como "livraria" para "livro". Em seguida, uma matriz TF-IDF é gerada para capturar a importância relativa dos termos nas transcrições; assim, as transcrições são agrupadas em *clusters* com base na semelhança dos termos (o número de *clusters* é definido pelo algoritmo K-Means). Para cada *cluster*, o LDA é aplicado para identificar os tópicos e palavras-chave relevantes. Dessa forma, quando uma pesquisa é realizada, a transcrição é atribuída a um *cluster*, e a busca ocorre dentro desse *cluster*, retornando uma lista de vídeos relevantes.

Huang e He (2024) apresentam uma abordagem para a tarefa de clusterização de textos, transformando-a em um problema de classificação com base em rótulos gerados pelas LLMs. Em vez de utilizar métodos tradicionais de processamento de linguagem natural (PLN) ou algoritmos de clusterização, como o K-Means, a técnica propõe o uso de LLMs para criar rótulos potenciais a partir dos dados textuais e classificar cada texto

nesses rótulos gerados. Essa abordagem simplifica o processo, eliminando a necessidade de ajustes complexos, reduzindo custos computacionais e aumentando a escalabilidade. Além disso, o método demonstrou desempenho superior em tarefas como mineração de tópicos, detecção de emoções e descoberta de intenções, destacando-se por sua eficiência, simplicidade e maior interpretabilidade dos agrupamentos formados.

Caron et al. (2018) propõem uma técnica chamada *DeepCluster*, voltada para a aprendizagem não supervisionada de representações visuais em imagens. O método combina *Convolutional Neural Network* (CNNs) com clusterização iterativa de características (*features*), utilizando principalmente o algoritmo K-Means para agrupar os dados. Esses agrupamentos geram pseudo-rótulos que são usados para treinar os pesos da rede em um ciclo iterativo de refinamento. Além do K-Means, os autores também exploram o *Power Iteration Clustering (PIC)* como alternativa, destacando que este não requer uma definição prévia do número de clusters, embora apresente maior desequilíbrio entre os tamanhos dos grupos formados. O *DeepCluster* se destaca pela sua escalabilidade e simplicidade, sendo capaz de lidar com conjuntos de dados massivos.

Onyepunuka et al. (2023) exploram como o algoritmo de recomendação do *YouTube* desvia usuários do tópico original para narrativas diferentes, usando a história de um almirante chinês do século XV como estudo de caso. Utilizando técnicas de PLN, o estudo analisou títulos de vídeos recomendados em cinco níveis de profundidade. A modelagem de tópicos (LDA) foi usada para identificar a mudança temática, medindo a distância de *Hellinger* entre as distribuições de tópicos das recomendações. Os resultados mostraram que, embora as recomendações adjacentes sejam semelhantes entre si, elas se afastam progressivamente do tópico inicial. A análise revelou um viés algorítmico no *YouTube* que prioriza vídeos com alta centralidade no grafo de recomendações, desviando a relevância temática ao longo das profundidades.

Todos os trabalhos citados utilizam um tipo de mídia, podendo ser vídeos, textos ou áudios, para conceber seus objetivos, assemelhando-se com este trabalho. Majoritariamente, a utilização do LDA para extração de tópicos em um grande volume de dados coincide com um dos métodos utilizados neste trabalho, usado para identificação de padrões semânticos para vídeos relacionados ao desenvolvimento de *software*. Embora a maioria dos artigos apresente formas de agrupamentos de vídeos, são focados em desenvolver ou apresentar um sistema novo que, baseado na clusterização, recomenda vídeos ou páginas *web*, diferente deste trabalho que visa melhorar os mecanismos de recomendação do *YouTube*. Além disso, este trabalho busca integrar técnicas de IA e modelagem de tópicos, com foco em otimizar a categorização de vídeos e melhorar as recomendações para os usuários do *YouTube*. Enquanto os trabalhos mencionados utilizam métodos tradicionais de PLN e clusterização como, LDA, K-Means, xMeans e *DenStream*, o presente trabalho propõe uma abordagem mais inovadora, que incorpora as LLMs para refinar a análise semântica dos vídeos, essa fusão de informações busca aprimorar a capacidade dos

algoritmos de recomendação em identificar padrões mais sutis e complexos, oferecendo uma experiência mais personalizada para os usuários. A Tabela 5 apresenta em suma a comparação dos trabalhos correlatos descritos neste capítulo com o trabalho proposto. É notório que a maioria dos artigos aborda clusterização e LDA aplicados a vídeos. Embora utilizem técnicas de clusterização e modelagem de tópicos, nenhum emprega LLMs como abordagem de clusterização. Este trabalho, por sua vez, propõe uma abordagem inovadora ao utilizar grandes modelos de linguagem para essa finalidade.

Tabela 5 – Comparação entre os trabalhos correlatos e o trabalho proposto

Autores	Tipos de dado	Técnica utilizada	Clusterização	Algoritmo
Basu, Yu e Zimmermann (2016)	Vídeos e textos (transcrições)	Modelagem de tópicos com LDA aplicada às transcrições dos vídeos para identificar tópicos latentes	Sim	LDA e Fuzzy C-Means
Basu et al. (2016)	Vídeos e textos (transcrições)	Modelagem de tópicos com LDA, recomendação de vídeos educativos com base em correspondência de tópicos	Não	LDA e Word2vec
Jansen et al. (2017)	Vídeos e áudios	Aprendizado não supervisionado e fraco supervisionado, detecção de atividade informativa, clusterização não paramétrica (*DenStream*) e algoritmos de *Locality-Sensitive Hashing* (LSH)	Sim	*DenStream*, LSH e PMI
Turcu et al. (2019)	Vídeos e textos (transcrições)	BoW, TF-IDF e LDA	Sim	K-Means, xMeans e LDA
Huang e He (2024)	Textos	Utiliza as LLMs para gerar rótulos e classificar	Sim	LLMs
Caron et al. (2018)	Imagens	CNNs para extração das imagens e clusterização não supervisionada para agrupar os recursos da rede	Sim	K-Means e *Power Iteration Clustering* (PIC)
Onyepunuka et al. (2023)	Vídeos	PLN e LDA para análise de redes usando métricas como modularidade e centralidade de autovetor	Sim	PLN e LDA
Este trabalho	Vídeos e textos (transcrições)	Modelagem de tópicos (LDA) e LLMs aplicada às transcrições dos vídeos para identificação de similaridade entre vídeos.	Sim	LLMs, LDA e LDADE

Capítulo 3

Experimentos e Análise dos Resultados

Este capítulo apresenta as metodologias empregadas para alcançar os objetivos desta monografia. Além da descrição dos procedimentos experimentais, são abordadas a análise e a discussão dos resultados obtidos.

3.1 Método para a Avaliação

Este trabalho adota um delineamento experimental de caráter exploratório, com o objetivo de comparar abordagens baseadas em grandes modelos de linguagem (LLMs) com métodos tradicionais de clusterização aplicados a conteúdos textuais derivados de vídeos.

O fluxograma apresentado tem como objetivo sumarizar as principais etapas do processo metodológico adotado neste trabalho. Inicialmente, realiza-se a seleção dos vídeos e é elaborada uma clusterização manual utilizada como referência para a avaliação dos agrupamentos gerados automaticamente, então, a obtenção de suas transcrições, que constituem a base de dados utilizada nos experimentos, em seguida, é conduzido o processo de clusterização das transcrições, utilizando tanto abordagens baseadas em LLMs quanto métodos tradicionais, como o LDA e sua variação otimizada, o LDADE. Por fim, os resultados obtidos são avaliados por meio de métricas externas de clusterização, possibilitando a análise comparativa entre os métodos e a verificação do alinhamento dos agrupamentos em relação à referência estabelecida.

Figura 4 – Etapas do método para a avaliação

Nas estratégias baseadas em LLMs, são consideradas duas abordagens principais para a clusterização das transcrições. Na primeira abordagem, transcrições de vídeos sobre desenvolvimento de *software*, extraídas da plataforma *YouTube*, são fornecidas diretamente a três LLMs (ChatGPT, Gemini e DeepSeek), que a partir do *prompt "A partir das transcrições acima, agrupe-as em cinco conjuntos por similaridade de conteúdo, os mais semelhantes ficam em um mesmo agrupamento. Cada transcrição deve pertencer a somente um conjunto, não podendo ser repetida em nenhum outro"* realizam o agrupamento das transcrições em *clusters* com base na similaridade semântica dos conteúdos.

A segunda abordagem baseia-se em um processo de duas etapas. Inicialmente, é fornecido às LLMs um *prompt* com a instrução: *"Gere um resumo conciso extraindo os 5 tópicos mais importantes do texto a seguir. Concentre-se em destacar informações relevantes que encapsulam os pontos principais do conteúdo."*, em seguida, a partir dos tópicos gerados, é utilizado o *prompt*: *"Com base nos tópicos gerados anteriormente, agrupe as transcrições no tópico que mais se assemelhe ao seu conteúdo. Cada transcrição deve pertencer somente a um único tópico, sem repetições."*, esse comando solicita a clusterização das transcrições com base na similaridade entre os conteúdos originais e os tópicos identificados.

Com o objetivo de avaliar os agrupamentos produzidos, os resultados das abordagens baseadas em LLMs são comparados com partições de referência, utilizadas como base comparativa. São consideradas três referências distintas, um modelo de *Latent Dirichlet Allocation* (LDA) com configurações padrão, utilizado como *baseline*; um modelo de LDA desenvolvido por Agrawal, Fu e Menzies (2018), no qual os hiperparâmetros são ajustados por meio de técnicas de otimização baseadas em DE; e uma clusterização manual construída por um avaliador humano, considerada como uma aproximação da interpretação semântica humana do conteúdo.

A comparação entre as partições é realizada por meio de métricas externas de avaliação de clusterização, especificamente o *Adjusted Rand Index* (ARI), o *Adjusted Mutual Information* (AMI), a *V-Measure* e o índice de Jaccard. Essas métricas permitem quantificar o grau de similaridade estrutural entre os agrupamentos gerados automaticamente e as referências adotadas.

Os dados utilizados no estudo consistem em transcrições de vídeos disponibilizadas pela plataforma *YouTube*. Foram selecionadas 15 transcrições relacionadas ao tema "Introdução à linguagem de programação Java", com duração entre 15 e 20 minutos. Para a identificação dos vídeos, utilizou-se o filtro de duração disponibilizado pela plataforma, conforme ilustrado na Figura 5, que permite selecionar conteúdos entre 4 e 20 minutos. A partir desse conjunto inicial, foram escolhidos apenas os vídeos com duração entre 15 e 20 minutos, de modo a garantir maior uniformidade no tamanho dos conteúdos analisados.

A escolha por vídeos com duração moderada também busca evitar conteúdos excessivamente longos, que tendem a abordar múltiplos tópicos em uma única transcrição, o que

poderia dificultar a definição de agrupamentos mais coesos, especialmente considerando que a clusterização de referência foi realizada por um único avaliador humano. Nesse contexto, conteúdos mais extensos poderiam aumentar a ambiguidade na atribuição dos clusters, elevando a probabilidade de inconsistências na classificação manual.

Para a formação dos agrupamentos, foi estabelecido o número de cinco clusters, considerando a quantidade de documentos e a necessidade de obter uma divisão equilibrada, que não fosse nem muito detalhada nem muito ampla, permitindo uma melhor distribuição das transcrições entre os grupos.

Todos os dados gerados no âmbito desta pesquisa encontram-se disponibilizados no repositório Zenodo[1]. O material inclui os links para os vídeos utilizados na etapa de avaliação, bem como suas respectivas transcrições. Também estão disponíveis os clusters gerados por cada uma das abordagens avaliadas, além dos resultados das métricas empregadas na análise. Adicionalmente, encontram-se acessíveis os algoritmos utilizados no desenvolvimento e na condução das análises realizadas neste estudo.

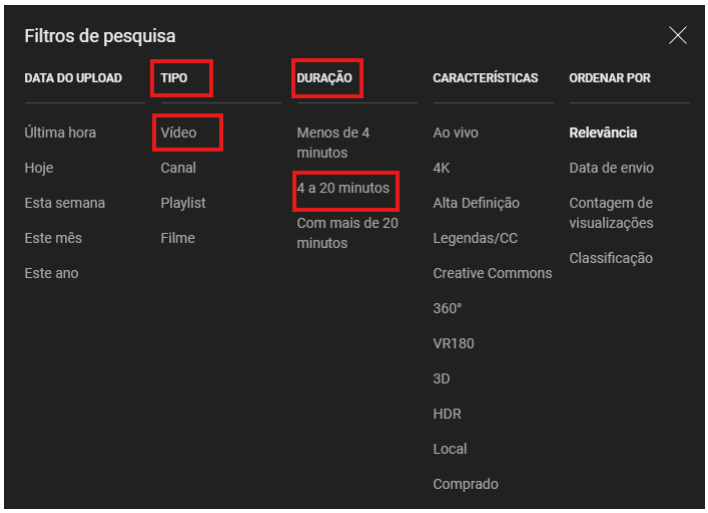

Figura 5 – Filtro do *YouTube*. Fonte: <https://www.youtube.com>

3.1.1 Métricas de avaliação de clusterização

A avaliação da qualidade de resultados obtidos por algoritmos de clusterização representa um desafio inerente às abordagens de aprendizado não supervisionado. Diferentemente de métodos supervisionados, nos quais rótulos verdadeiros estão disponíveis para validação direta, a clusterização não dispõe, em geral, de uma resposta correta previamente definida. Dessa forma, torna-se necessário o uso de métricas específicas capazes de quantificar o grau de similaridade, concordância ou consistência entre diferentes particionamentos dos dados.

[1] https://zenodo.org/records/18890800

As métricas de avaliação de clusterização podem ser classificadas, de modo geral, em métricas internas e métricas externas. As métricas internas avaliam a qualidade dos agrupamentos com base apenas nas informações intrínsecas aos dados, considerando aspectos como coesão *intra-cluster* e separação *inter-cluster*. Já as métricas externas comparam os clusters obtidos com uma partição de referência, permitindo mensurar o quão semelhante é o agrupamento gerado em relação a uma classificação previamente conhecida ou a outro método de clusterização.

3.1.1.1 ARI

O ARI (*Adjusted Rand Index*) é uma métrica amplamente utilizada para avaliar a similaridade entre dois agrupamentos, levando em consideração a correção para concordâncias que podem ocorrer ao acaso. Diferentemente de métricas não ajustadas, o ARI fornece uma medida mais robusta, pois desconta a similaridade esperada entre partições aleatórias, permitindo uma interpretação mais confiável dos resultados (HUBERT; ARABIE, 1985).

O ARI é uma versão ajustada do *Rand Index* (RI), que mede a concordância entre dois agrupamentos com base em todos os pares possíveis de amostras do conjunto de dados. Para cada par de amostras, avalia-se se elas pertencem ao mesmo *cluster* ou a *clusters* distintos em ambas as partições. A métrica contabiliza como concordantes os pares que estão juntos em ambos os agrupamentos ou separados em ambos, e como discordantes os pares que apresentam classificações diferentes entre as partições.

Entretanto, o RI não considera a possibilidade de concordâncias ocorrendo de forma aleatória, o que pode levar a valores inflados, especialmente em cenários com muitos *clusters* ou distribuições desequilibradas. Para contornar essa limitação, o ARI introduz uma correção estatística baseada no valor esperado do RI sob um modelo de aleatoriedade, no qual as atribuições de *clusters* são feitas ao acaso, preservando apenas o tamanho dos grupos.

Formalmente, o ARI é definido como:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]}$$

Essa normalização garante que o ARI assuma valores no intervalo $[-1, 1]$. Um valor igual a 1 indica concordância perfeita entre os agrupamentos, valores próximos de 0 sugerem que a similaridade observada é equivalente à esperada ao acaso, enquanto valores negativos indicam concordância inferior à aleatória. Dessa forma, o ARI torna-se especialmente adequado para comparar resultados de algoritmos de clusterização com uma partição de referência ou entre diferentes abordagens de agrupamento.

Como a métrica é baseada na comparação par a par entre amostras (pares que permanecem juntos ou separados nas duas partições), o ARI não depende de qual agrupamento

é tratado como referência. Assim, vale a propriedade de simetria:

$$ARI(a, b) = ARI(b, a)$$

Em síntese, o ARI destaca-se como uma métrica robusta para a avaliação da similaridade entre agrupamentos, pois incorpora uma correção estatística que elimina o efeito de concordâncias aleatórias. Ao contrário do RI, que pode apresentar valores elevados mesmo na ausência de estrutura real nos dados, o ARI fornece uma medida normalizada e mais fiel da qualidade dos resultados de clusterização.

Além disso, propriedades como a simetria e a interpretação bem definida de seus valores variando de -1 a 1 tornam o ARI particularmente adequado para a comparação entre diferentes algoritmos de agrupamento ou entre uma partição gerada e uma partição de referência.

3.1.1.2 AMI

O AMI (*Adjusted Mutual Information*) é uma métrica utilizada para medir a similaridade entre dois agrupamentos com base na teoria da informação, avaliando a quantidade de informação compartilhada entre duas partições. Assim como o ARI, o AMI incorpora uma correção para o acaso, tornando a comparação entre *clusters* mais robusta e confiável (scikit-learn developers, 2024).

O AMI é uma versão ajustada da *Mutual Information* (MI) que quantifica a dependência estatística entre dois agrupamentos ao medir quanto o conhecimento de uma partição reduz a incerteza sobre a outra. Em termos práticos, a métrica avalia o quanto as atribuições de *clusters* em uma partição informam sobre as atribuições na outra.

Entretanto, a MI apresenta uma limitação importante: seus valores tendem a aumentar conforme cresce o número de *clusters*, mesmo quando os agrupamentos são gerados aleatoriamente. Esse comportamento pode levar a interpretações incorretas da similaridade entre partições. Para mitigar esse efeito, o AMI ajusta a informação mútua observada subtraindo seu valor esperado sob um modelo de aleatoriedade, no qual as partições são consideradas independentes, mantendo apenas a distribuição dos tamanhos dos *clusters*.

Formalmente, o AMI é definido como:

$$AMI = \frac{MI - \mathbb{E}[MI]}{\max(MI) - \mathbb{E}[MI]}$$

Essa normalização garante que o AMI assuma valores no intervalo $[0, 1]$, onde valores próximos de 1 indicam alta concordância entre os agrupamentos, enquanto valores próximos de 0 indicam similaridade equivalente à esperada ao acaso. Dessa forma, o AMI fornece uma medida interpretável e comparável entre diferentes cenários de clusterização.

Além disso, o AMI é simétrico em relação às partições comparadas, ou seja, a ordem dos agrupamentos não influencia o resultado da métrica:

$$AMI(a, b) = AMI(b, a)$$

Portanto, o AMI constitui uma métrica adequada para a avaliação da similaridade entre agrupamentos quando se deseja capturar relações baseadas em dependência estatística entre partições. Ao corrigir a informação mútua para o efeito do acaso, o AMI evita tendências associados ao número de *clusters* e à distribuição de seus tamanhos, proporcionando uma análise mais justa e consistente.

Suas propriedades de normalização, simetria e interpretação clara tornam o AMI especialmente útil para a comparação entre diferentes algoritmos de clusterização ou entre agrupamentos gerados a partir de abordagens distintas.

3.1.1.3 *V-Measure*

A *V-Measure* é uma métrica externa de avaliação de clusterização baseada na teoria da informação, proposta para medir a qualidade de um agrupamento a partir de dois critérios complementares: homogeneidade e completude (ROSENBERG; HIRSCHBERG, 2007). Essa métrica avalia o quão bem os *clusters* formados refletem uma partição de referência, considerando simultaneamente a pureza dos grupos e a consistência na atribuição dos elementos.

A homogeneidade mensura se cada *cluster* contém predominantemente elementos pertencentes a uma única classe de referência, enquanto a completude avalia se todos os elementos de uma mesma classe estão concentrados majoritariamente em um único *cluster*. Esses dois critérios são expressos em termos de entropia condicional, permitindo uma interpretação probabilística da qualidade do agrupamento.

Entretanto, a avaliação isolada de apenas um desses critérios pode levar a interpretações incompletas. Um agrupamento pode apresentar alta homogeneidade ao criar muitos *clusters* pequenos, mas baixa completude, ou vice-versa. Para equilibrar esses aspectos, a *V-Measure* combina homogeneidade e completude por meio da média harmônica, assegurando que ambos contribuam de forma equilibrada para o valor final da métrica.

Formalmente, a *V-Measure* é definida como:

$$V = \frac{(1 + \beta) \cdot h \cdot c}{\beta \cdot h + c}$$

em que h representa a homogeneidade, c a completude e β é um parâmetro que permite ponderar a importância relativa entre esses dois critérios. Quando $\beta = 1$, homogeneidade e completude possuem pesos iguais.

A normalização da métrica garante que a *V-Measure* assuma valores no intervalo $[0, 1]$, onde valores próximos de 1 indicam alta correspondência entre os agrupamentos e a partição de referência, enquanto valores próximos de 0 indicam baixa concordância. Além disso, a *V-Measure* é simétrica em relação às partições comparadas, não sendo afetada pela ordem dos agrupamentos analisados.

Dessa forma, a *V-Measure* constitui uma métrica adequada para a avaliação da qualidade de resultados de clusterização quando se deseja capturar simultaneamente a pureza dos grupos e a consistência na atribuição das classes. Sua interpretação clara e balanceada torna essa métrica especialmente útil na comparação entre diferentes algoritmos de clusterização.

3.1.1.4 Índice de Jaccard

O índice de Jaccard é uma métrica amplamente utilizada para medir a similaridade entre dois conjuntos, sendo aplicado, no contexto da clusterização, para avaliar a concordância entre dois agrupamentos a partir da comparação par a par dos elementos. Essa métrica quantifica a proporção de pares de elementos que são agrupados conjuntamente em ambas as partições em relação ao total de pares que aparecem juntos em pelo menos uma delas (JACCARD, 1901).

No cenário de avaliação de clusterização, o índice de Jaccard considera apenas as associações positivas entre elementos, isto é, os pares que pertencem ao mesmo *cluster* em pelo menos uma das partições analisadas. Dessa forma, a métrica ignora pares que estão separados em ambos os agrupamentos, concentrando-se exclusivamente na coincidência das atribuições de *clusters*.

Formalmente, o índice de Jaccard pode ser definido como:

$$J = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

em que M_{11} representa o número de pares de elementos que pertencem ao mesmo *cluster* em ambas as partições, M_{10} o número de pares agrupados conjuntamente apenas na primeira partição e M_{01} o número de pares agrupados conjuntamente apenas na segunda partição.

A normalização da métrica garante que o índice de Jaccard assuma valores no intervalo $[0, 1]$, onde valores próximos de 1 indicam alta similaridade entre os agrupamentos, enquanto valores próximos de 0 indicam baixa concordância. Além disso, a métrica é simétrica em relação às partições comparadas, ou seja, a ordem dos agrupamentos não influencia o valor obtido:

$$J(a, b) = J(b, a)$$

Entretanto, diferentemente de métricas ajustadas como o ARI e o AMI, o índice de Jaccard não incorpora correção para o acaso. Como consequência, seus valores podem ser influenciados pela quantidade de *clusters* e pela distribuição de seus tamanhos, especialmente em cenários com grande número de grupos ou partições desbalanceadas.

Apesar dessa limitação, o índice de Jaccard permanece uma métrica simples, interpretável e computacionalmente eficiente para a comparação de agrupamentos, sendo particularmente útil como medida complementar em análises de clusterização. Quando utilizado

em conjunto com métricas ajustadas, o índice de Jaccard contribui para uma avaliação mais abrangente da similaridade entre diferentes particionamentos dos dados.

3.2 Experimentos

Os experimentos foram conduzidos com o objetivo de avaliar comparativamente o desempenho das abordagens baseadas em LLMs e dos métodos de modelagem de tópicos na tarefa de clusterização semântica das transcrições dos vídeos realicionados ao desenvolvimento de *software* do *YouTube*. O conjunto de dados utilizado corresponde ao descrito na Seção 3.1.

3.2.1 Pipeline Experimental

O procedimento experimental envolveu a preparação dos dados, a execução das abordagens baseadas em modelagem de tópicos, a aplicação das estratégias baseadas em LLMs e, por fim, a avaliação comparativa das partições obtidas.

As transcrições dos vídeos foram submetidas a pré-processamento textual, incluindo normalização, remoção de *stop words* e lematização. Esse processamento foi aplicado às abordagens baseadas em LDA, enquanto as LLMs operaram diretamente sobre o texto original.

3.2.2 Modelagem de Tópicos

Foram avaliadas duas configurações do modelo LDA, a versão padrão, com hiperparâmetros definidos manualmente, e o LDADE, com ajuste automático por meio de DE, conforme implementação disponibilizada por Agrawal, Fu e Menzies (2018).

Após o treinamento, cada transcrição foi associada ao tópico de maior probabilidade na distribuição temática inferida pelo modelo, resultando em uma atribuição exclusiva de rótulo por documento.

3.2.3 Abordagens Baseadas em LLMs

Foram executadas as duas estratégias descritas na Seção 3.1. Os modelos avaliados diferem quanto à arquitetura e ao tipo de acesso (gratuito ou pago), conforme apresentado na Tabela 6.

Cada modelo produziu uma partição contendo cinco agrupamentos, respeitando a restrição de atribuição exclusiva de cada transcrição a um único grupo.

Tabela 6 – Modelos de LLMs utilizados nos experimentos

Modelo	**Versão**	**Tipo de acesso**
ChatGPT	GPT-4.1	Pago
ChatGPT	GPT-4o mini	Gratuito
Gemini	2.5 Flash	Gratuito
DeepSeek	DeepSeek-V3	Gratuito

3.2.4 Referência Humana

Uma clusterização manual foi realizada por um avaliador humano, que analisou as transcrições individualmente e as agrupou segundo critérios de similaridade temática. Essa partição foi utilizada como referência adicional na comparação com as partições produzidas pelas abordagens baseadas em LLMs e pelos modelos de LDA.

3.2.5 Avaliação

As partições produzidas por todas as abordagens foram comparadas por meio das métricas externas de avaliação de clusterização *Adjusted Rand Index* (ARI), *Adjusted Mutual Information* (AMI), *V-Measure* e índice de Jaccard, conforme definidas na Seção 3.1.1, permitindo mensurar o grau de concordância estrutural entre os agrupamentos gerados. A análise concentra-se na comparação quantitativa entre métodos, considerando o caráter exploratório do estudo e o tamanho reduzido do conjunto de dados.

3.3 Resultados

Esta seção apresenta os resultados obtidos a partir da comparação entre diferentes abordagens de clusterização aplicadas a transcrições de vídeos, incluindo abordagens tradicionais baseadas em modelagem em tópicos e modelos de LLMs. O foco da análise é avaliar o desempenho de clusterização semântica dos diferentes métodos, medido pelo grau de similaridade entre os agrupamentos gerados automaticamente e uma clusterização manual de referência, adotada como aproximação da interpretação temática humana do conteúdo. Para isso, foram utilizadas métricas externas de avaliação, permitindo uma comparação quantitativa e estrutural entre as partições produzidas.

A análise dos resultados evidencia que não há uma similaridade global elevada entre todos os métodos avaliados. Contudo, observa-se de forma consistente a formação de subconjuntos de métodos com alta concordância interna. Esses subconjuntos concentram-se predominantemente nos modelos baseados em LLMs com maior capacidade de contexto, enquanto métodos tradicionais e versões com recursos reduzidos apresentam menor alinhamento tanto entre si quanto em relação às referências adotadas. Esse comportamento é observado de maneira recorrente nas Figuras 6, 7, 8 e 9, correspondentes às métricas ARI, AMI, *V-Measure* e índice de Jaccard, respectivamente.

Para facilitar a análise comparativa entre os métodos avaliados, os resultados foram representados por meio de *heatmaps*. Um *heatmap* é uma representação visual matricial na qual os valores numéricos de similaridade são codificados por meio de uma escala de cores, permitindo identificar de forma intuitiva padrões, concentrações e discrepâncias entre os métodos. Cores mais intensas indicam maior grau de similaridade, enquanto cores menos intensas representam menor concordância. Essa forma de visualização é particularmente adequada para métricas par a par, pois possibilita identificar grupos de métodos com comportamento semelhante e comparar diretamente seu alinhamento com a clusterização manual de referência.

Cabe destacar que todas as matrizes de similaridade apresentadas nos *heatmaps* que serão apresentados a seguir, são simétricas em relação à diagonal principal. Essa simetria decorre da definição das métricas externas utilizadas, que satisfazem a propriedade de simetria, isto é, a similaridade entre dois métodos é independente da ordem de comparação. Assim, os *heatmaps* refletem visualmente essa propriedade matemática, reforçando a consistência dos resultados obtidos.

No *heatmap* apresentado na Figura 6, referente ao *Adjusted Rand Index* (ARI), observa-se que os modelos com maior capacidade de contexto, como o ChatGPT na versão paga e o DeepSeek, apresentam valores mais elevados de similaridade entre si, além de maior proximidade em relação à clusterização manual. Nos gráficos, a sigla "abd" corresponde a abordagem, indicando as duas estratégias experimentais baseadas em LLMs adotadas neste estudo, a Abordagem 1 (abd 1), na qual as transcrições foram fornecidas diretamente aos modelos para agrupamento semântico, e a Abordagem 2 (abd 2), baseada na geração intermediária de representações estruturadas antes da clusterização.

A análise evidencia que, no caso do Gemini, a Abordagem 1 apresentou desempenho superior em comparação à Abordagem 2, indicando que, para esse modelo, a clusterização direta das transcrições produziu agrupamentos mais consistentes do que a estratégia baseada em representações intermediárias. Em contrapartida, ao comparar os pares de abordagens entre os modelos, observa-se que o ChatGPT em sua versão paga na Abordagem 2 e o DeepSeek na Abordagem 2 apresentam os maiores níveis de alinhamento entre si, indicando maior simetria estrutural na formação dos *clusters*.

Dado que o ARI é uma métrica ajustada ao acaso, os resultados indicam que a concordância observada não decorre de agrupamentos aleatórios, mas reflete uma organização semântica mais consistente. Além disso, os métodos baseados em modelagem de tópicos demonstram sensibilidade à parametrização. Em comparação com o agrupamento manual, o LDADE apresentou desempenho superior ao LDA tradicional em configuração padrão, evidenciando que o ajuste de hiperparâmetros influencia diretamente a qualidade dos agrupamentos obtidos. Ainda assim, mesmo com otimização, os métodos probabilísticos apresentaram maior variabilidade quando comparados aos modelos baseados em LLMs.

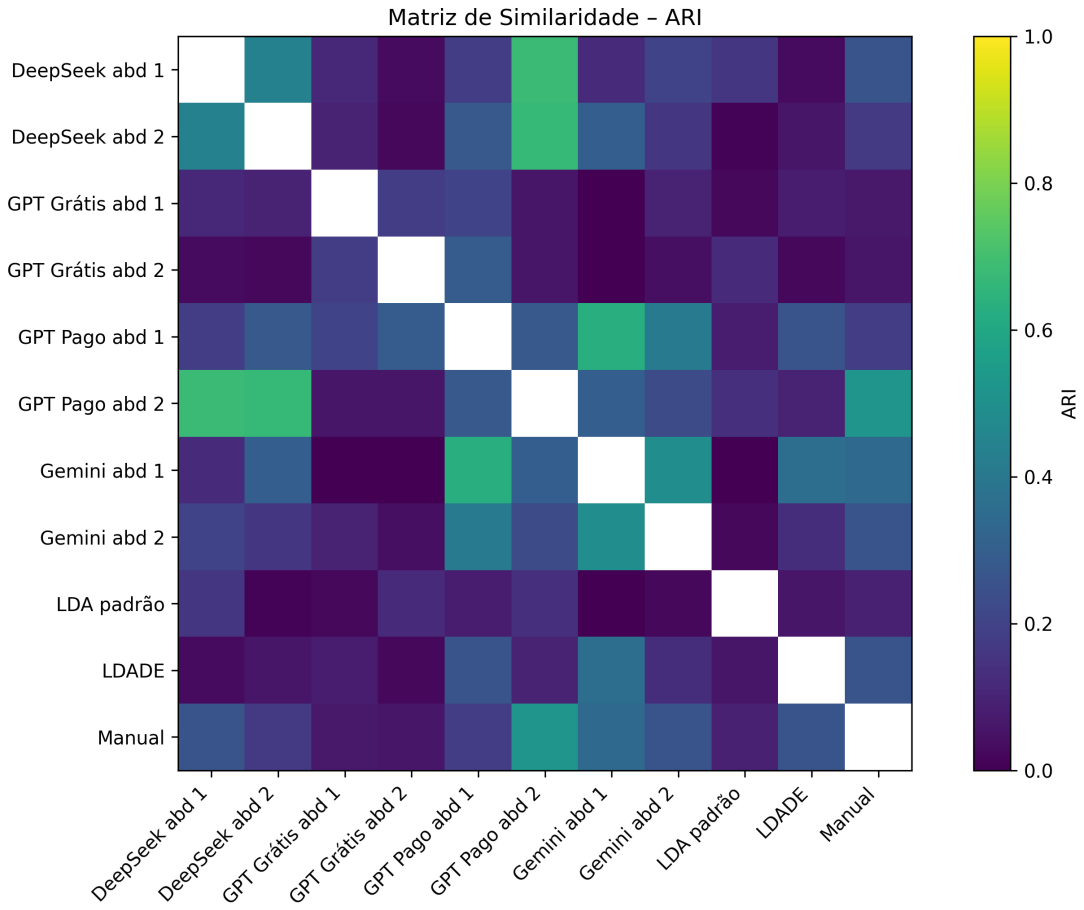

Figura 6 – Heatmap de Similaridade ARI

A Figura 7, referente ao *Adjusted Mutual Information* (AMI), reforça o padrão identificado anteriormente, sob a perspectiva do compartilhamento de informação entre partições. Observa-se que o Gemini apresentou desempenho superior na Abordagem 1 em relação à Abordagem 2, indicando que a clusterização direta preservou maior quantidade de informação mútua quando comparada à clusterização manual de referência.

Em relação aos pares de abordagens, o ChatGPT em sua versão paga na Abordagem 2 e o DeepSeek na Abordagem 2 mantêm os maiores níveis de alinhamento entre si, evidenciando elevada similaridade estrutural também sob a métrica de informação mútua ajustada ao acaso adotada pelo AMI.

No contexto da modelagem de tópicos, o LDADE novamente superou o LDA tradicional em configuração padrão quando comparado ao agrupamento manual, confirmando que a calibração de hiperparâmetros exerce impacto direto na capacidade de retenção das relações latentes presentes nos dados textuais.

Em relação à *V-Measure*, ilustrada na Figura 8, observa-se uma elevação geral dos valores quando comparada às métricas ARI e AMI. Esse comportamento é consistente com a natureza da métrica, que combina homogeneidade e completude por meio da média harmônica, tornando-a menos sensível a pequenas divergências estruturais entre *clusters*.

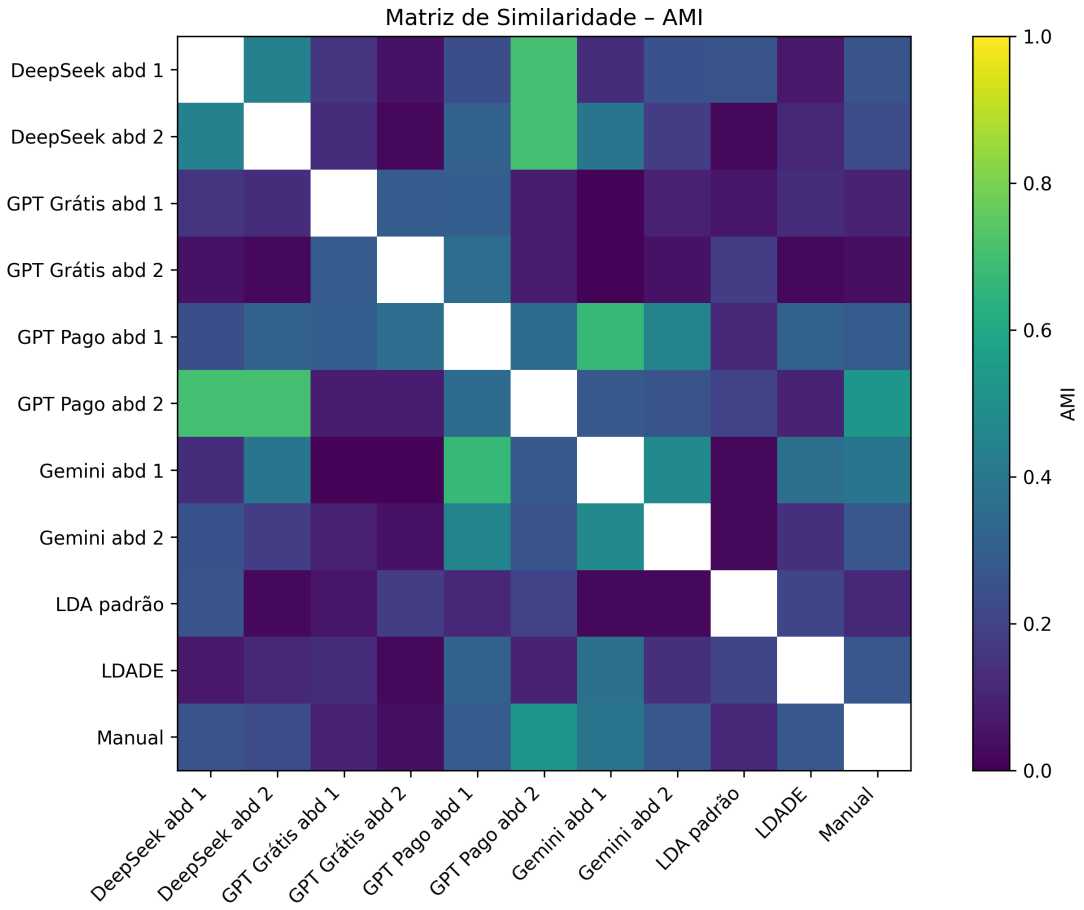

Figura 7 – Heatmap de Similaridade AMI

Mesmo sob essa suavização, o Gemini manteve desempenho superior na Abordagem 1 em relação à Abordagem 2, indicando maior equilíbrio entre a pureza interna dos *clusters* e a capacidade de reunir elementos semanticamente relacionados em um mesmo agrupamento. De forma convergente ao observado anteriormente, o ChatGPT em sua versão paga na Abordagem 2 e o DeepSeek na Abordagem 2 continuam a apresentar os maiores níveis de alinhamento entre si, evidenciando consistência também nos critérios de homogeneidade e completude.

No âmbito da modelagem de tópicos, o LDADE novamente superou o LDA tradicional em configuração padrão na maioria das comparações, sugerindo que a otimização do *pipeline* contribui para a formação de agrupamentos mais equilibrados. Ainda assim, os modelos baseados em LLMs demonstraram maior estabilidade global na organização dos *clusters*. Dessa forma, a *V-Measure* confirma que os modelos com maior capacidade semântica tendem a produzir partições mais consistentes sob múltiplos critérios de avaliação.

Por fim, o índice de Jaccard, apresentado na Figura 9, preserva a mesma tendência relativa entre os métodos, embora com valores absolutos mais baixos. Como essa métrica considera exclusivamente a proporção de pares de elementos agrupados conjuntamente

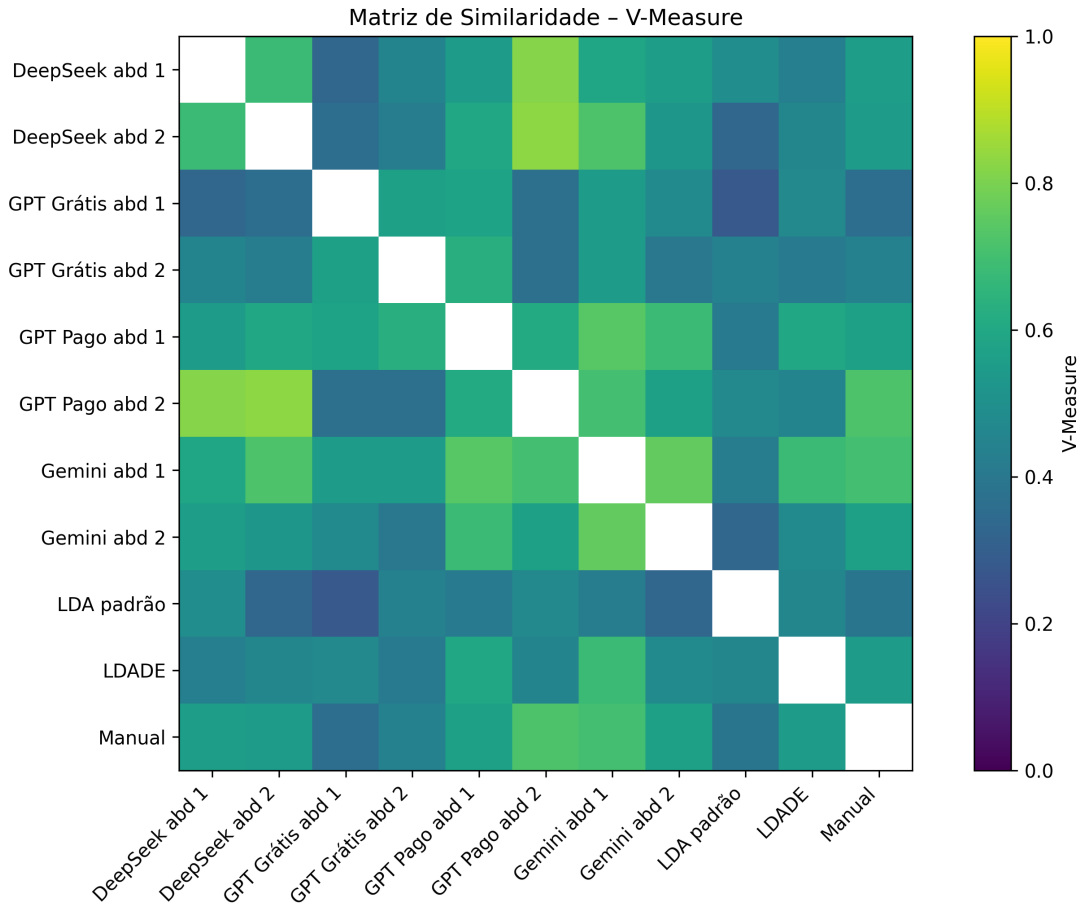

Figura 8 – Heatmap de Similaridade V-Measure

em ambas as partições, sem ajuste para o acaso, torna-se mais sensível à fragmentação excessiva ou à formação de *clusters* muito pequenos.

Mesmo sob esse critério mais restritivo, o Gemini apresentou desempenho superior na Abordagem 1 em comparação à Abordagem 2 na maioria dos casos, indicando maior coincidência direta entre os pares de documentos agrupados conjuntamente. De forma consistente com as métricas anteriores, o ChatGPT em sua versão paga na Abordagem 2 e o DeepSeek na Abordagem 2 mantêm o maior grau de similaridade entre si, evidenciando alinhamento estrutural também quando analisados apenas sob a interseção de pares coincidentes.

No que se refere à modelagem de tópicos, o LDADE novamente superou o LDA tradicional quando foram comparados ao agrupamento manual, ainda que ambos tenham sido mais impactados pela sensibilidade do índice à fragmentação dos agrupamentos. Assim, os resultados obtidos com Jaccard reforçam o padrão geral observado, embora sob uma perspectiva mais conservadora de similaridade. Observa-se, portanto, que pequenas divergências na composição interna dos *clusters* tendem a produzir reduções mais acentuadas nessa métrica. Ainda assim, mesmo usando métricas diferentes, os métodos mantêm a mesma posição relativa, o que fortalece a confiabilidade dos resultados.

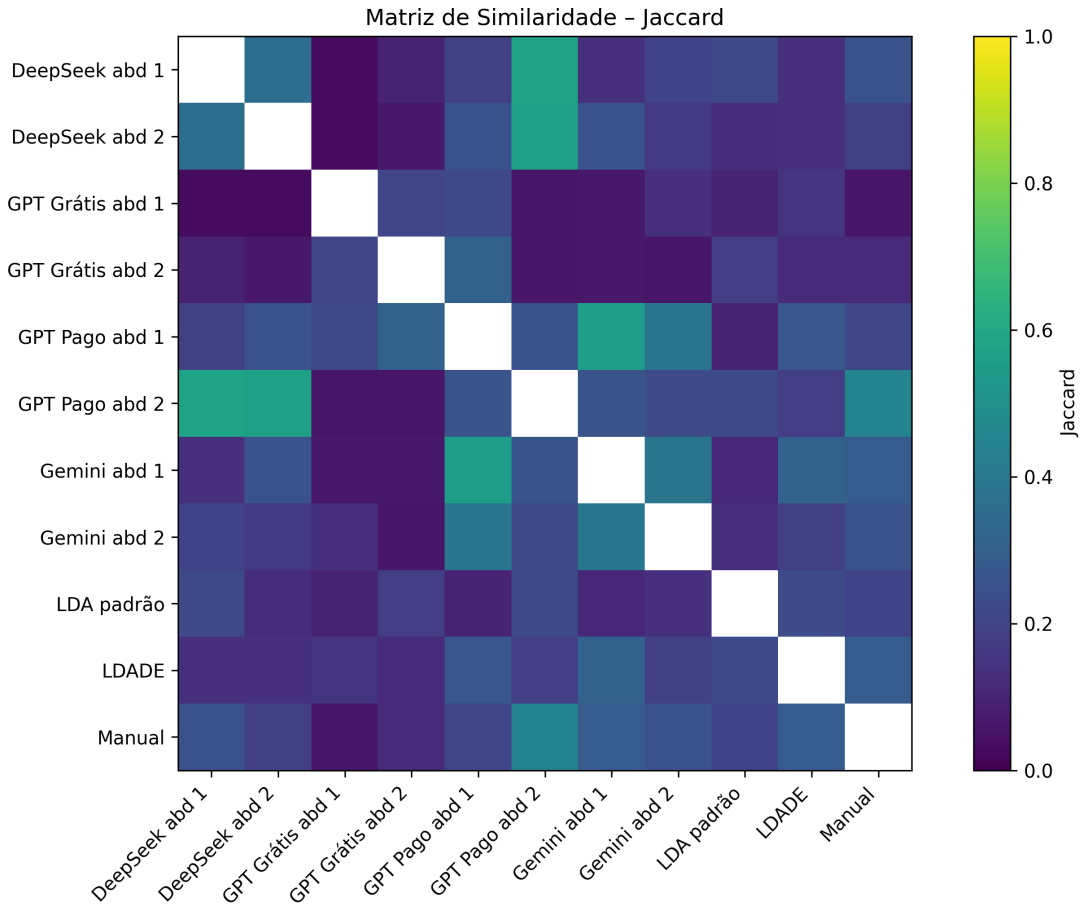

Figura 9 – Heatmap de Similaridade Jaccard

Cabe ressaltar que métodos tradicionais de clusterização baseados em modelagem de tópicos continuam sendo abordagens válidas e interpretáveis, especialmente quando bem configurados (LDADE). Os resultados observados indicam que modelos baseados em LLMs tendem a oferecer maior robustez semântica, enquanto métodos clássicos demonstram desempenho dependente de parametrização e pré-processamento.

De modo geral, os resultados indicam que modelos baseados em LLMs, especialmente em configurações com maior capacidade, produzem agrupamentos mais consistentes, estruturalmente estáveis e semanticamente alinhados com a referência humana quando comparados a abordagens com menor capacidade contextual.

3.4 Discussão

Os resultados obtidos permitem analisar o comportamento dos diferentes métodos de clusterização sob a perspectiva da qualidade estrutural dos agrupamentos e de seu alinhamento semântico com a referência humana. Considerando o objetivo deste trabalho descrito na Seção 1.4, observa-se que os modelos com maiores capacidades semânticas, como o ChatGPT em sua versão paga e o DeepSeek, apresentaram de forma consistente,

maiores níveis de similaridade com a clusterização manual e maior concordância interna entre si, especialmente nas configurações com maior capacidade de contexto.

Esse comportamento sugere que modelos de linguagem de grande porte são capazes de produzir representações textuais mais semanticamente estruturadas, o que impacta positivamente a etapa de agrupamento. Diferentemente de métodos puramente estatísticos baseados em frequência de termos (LDA), as LLMs operam com representações distribuídas e contextualizadas, permitindo capturar relações conceituais que não dependem exclusivamente da coocorrência direta de palavras. Essa característica favorece a formação de *clusters* tematicamente mais coerentes, sobretudo em textos longos e semanticamente ricos, como transcrições de vídeos.

Aspectos arquiteturais ajudam a interpretar esse resultado. Conforme apresentado na Tabela 3, parte dos modelos analisados utiliza variações da arquitetura *Transformer*, com possíveis extensões como MoE. Embora nem todos os fornecedores divulguem detalhes completos de implementação, a literatura indica que esses mecanismos ampliam a capacidade representacional e a especialização interna do modelo. Como consequência, a geração de resumos e descrições temáticas tendem a preservar melhor o conteúdo semântico central dos documentos.

O mecanismo de *Multi-Token Prediction* (MTP) também contribui para essa interpretação. Ao permitir a predição de múltiplos tokens de forma integrada, o modelo considera dependências contextuais mais amplas durante a geração textual. Isso favorece saídas mais coerentes e semanticamente estáveis, o que impacta indiretamente a qualidade das representações utilizadas na clusterização. Dessa forma, a etapa de geração mediada por LLMs não apenas resume o conteúdo, mas também organiza semanticamente as informações de modo mais consistente.

Os métodos tradicionais baseados em modelagem de tópicos, como o LDA, também demonstraram capacidade de formar agrupamentos relevantes, especialmente quando submetidos a ajustes de parâmetros e pré-processamento adequado (LDADE). Os resultados indicam que o desempenho não depende só do modelo, depende muito de como os dados são preparados. Isso é esperado, dado que tais métodos operam majoritariamente com representações do tipo *Bag-of-Words* (BoW), sem incorporar diretamente dependências contextuais profundas. Assim, não se trata de inadequação do método, mas de diferenças de natureza representacional entre as abordagens.

Observou-se ainda diferença entre configurações de modelos de linguagem no contexto experimental deste estudo. Versões mais robustas das LLMs, como o ChatGPT em sua versão paga e o DeepSeek, apresentaram valores consistentemente superiores nas métricas externas de avaliação, bem como maior alinhamento com a clusterização manual, quando comparadas ao ChatGPT em sua versão gratuita e ao Gemini. Essa superioridade manifestou-se também em maior estabilidade estrutural entre as abordagens analisadas. Tais resultados indicam que a forma como os modelos processam e organizam semantica-

mente o conteúdo influencia o desempenho observado nas métricas de clusterização.

A interpretação dos resultados também deve considerar o comportamento das métricas utilizadas. Medidas ajustadas ao acaso, como ARI e AMI, mostraram-se mais estáveis para comparação estrutural e informacional das partições. A *V-Measure*, por sua vez, apresentou valores globalmente mais elevados, o que é esperado em função de sua formulação baseada na média harmônica entre homogeneidade e completude, tornando-a menos sensível a pequenas divergências estruturais. Já o índice de Jaccard apresentou valores mais conservadores, o que é compatível com sua maior sensibilidade à fragmentação, à presença de *clusters* pequenos e a *clusters* vazios. Por essa razão, a análise das métricas deve ser realizada de forma complementar, evitando interpretações isoladas.

Considerando os resultados obtidos, a hipótese de que modelos baseados em LLMs são capazes de produzir clusterizações semanticamente consistentes é sustentada pelos dados experimentais, dentro das condições avaliadas. As evidências indicam que essas abordagens constituem alternativas viáveis para tarefas de organização temática automática.

Capítulo 4

Conclusão

Este trabalho teve como objetivo investigar a eficácia da utilização de grandes modelos de linguagem (LLMs) em comparação com métodos tradicionais baseados em modelagem de tópicos no agrupamento semântico de transcrições de vídeos educacionais relacionados ao desenvolvimento de *software*. Ao longo do desenvolvimento da pesquisa, foram exploradas diferentes abordagens experimentais, métricas de avaliação e referências comparativas, permitindo analisar de forma sistemática o comportamento dessas técnicas em um cenário controlado.

O objetivo geral de comparar abordagens baseadas em LLMs com métodos tradicionais baseados em modelagem de tópicos (LDA) foi atingido por meio da aplicação de múltiplas estratégias de clusterização e da utilização de métricas externas de avaliação, como ARI, AMI, *V-Measure* e índice de Jaccard, bem como da comparação com uma clusterização manual de referência. Os objetivos específicos, que incluíram a análise do impacto da capacidade contextual dos modelos, a investigação do desempenho de versões gratuitas e pagas de LLMs e a avaliação da sensibilidade dos métodos tradicionais à parametrização, também foram contemplados ao longo dos experimentos e da análise dos resultados.

Os resultados obtidos permitiram concluir que o desenvolvimento metodológico adotado foi adequado para responder à questão de pesquisa proposta, fornecendo evidências empíricas que sustentam que abordagens baseadas em grandes modelos de linguagem (LLMs) produzem agrupamentos semanticamente mais consistentes de transcrições de vídeos quando comparadas a métodos tradicionais de clusterização baseados em modelagem de tópicos.

Através dos dados apurados por meio das métricas, especialmente ARI e AMI, indicaram maior concordância entre os agrupamentos gerados por LLMs e a clusterização manual, reforçando a hipótese inicial de que representações textuais contextualizadas favorecem a captura de relações semânticas mais profundas. A *V-Measure* apresentou valores globalmente mais elevados, evidenciando maior equilíbrio entre homogeneidade e completude dos agrupamentos, o que confirma a consistência estrutural dos resultados sob diferentes critérios de avaliação. Além disso, observou-se que métodos tradicionais, como

a modelagem de tópicos (LDA), continuam sendo alternativas válidas, desde que acompanhadas de pré-processamento adequado e ajustes criteriosos de parâmetros, embora apresentem maior sensibilidade a essas configurações.

Além disso, destaca-se como principal diferencial desta pesquisa a utilização de grandes modelos de linguagem (LLMs) não apenas como técnica de clusterização, mas também como uma forma de referência semântica para analisar os agrupamentos gerados por métodos tradicionais, como LDA e LDADE. Essa abordagem permite uma comparação mais alinhada à interpretação de similaridade de conteúdo, aproximando a análise automática de uma perspectiva mais contextual e próxima da compreensão humana.

Dessa forma, os resultados obtidos validam a hipótese de pesquisa e demonstram a viabilidade do uso de LLMs como alternativas aos métodos tradicionais de clusterização textual.

Cabe ainda destacar algumas limitações deste estudo, o conjunto de dados utilizado é relativamente pequeno e restrito a um único domínio temático, o que limita a generalização dos resultados. A clusterização manual de referência foi realizada por um único avaliador, podendo incorporar vieses subjetivos. Além disso, saídas de LLMs podem apresentar variabilidade associada a *prompt* e configuração de geração. Estudos futuros com conjuntos maiores, múltiplos avaliadores humanos e controle de variabilidade de geração podem aprofundar a análise aqui apresentada.

4.1 Principais Contribuições

Este trabalho apresenta as seguintes contribuições:

(1) A principal contribuição deste trabalho consiste na análise comparativa sistemática entre métodos tradicionais de clusterização baseados em modelagem de tópicos e abordagens baseadas em grandes modelos de linguagem aplicadas ao agrupamento semântico de transcrições de vídeos educacionais. Essa contribuição está associada à proposição de um cenário experimental controlado que permite analisar o comportamento dessas abordagens sob diferentes critérios de avaliação.

(2) Outra contribuição relevante é a inclusão de uma referência humana como aproximação da interpretação semântica, permitindo uma avaliação mais próxima da percepção humana do conteúdo. Essa abordagem fortalece a análise comparativa e evidencia o potencial das LLMs como ferramentas auxiliares em tarefas de organização temática automática, por meio de transcrições de vídeos.

(3) Nesse contexto, do ponto de vista aplicado, os resultados sugerem potencial de uso em sistemas de organização de transcrições textuais, indexação temática e apoio à categorização automática de conteúdos derivados de vídeos. A capacidade de gerar representações semanticamente ricas pode reduzir a dependência de rotulação manual extensiva, servindo como etapa de apoio na organização de grandes coleções de textos.

4.2 Trabalhos Futuros

Apesar dos resultados promissores, este trabalho apresenta limitações que abrem espaço para investigações futuras. Uma das melhorias possíveis é ampliar o conjunto de dados, tanto em quantidade quanto em diversidade temática, permitindo avaliar a generalização dos resultados para outros domínios além do desenvolvimento de *software*. A inclusão de transcrições em diferentes idiomas ou provenientes de distintos contextos educacionais também pode contribuir para análises mais abrangentes.

Outra extensão relevante envolve a utilização de múltiplos avaliadores humanos na construção das partições de referência, reduzindo possíveis vieses subjetivos e permitindo o cálculo de concordância entre avaliadores. Além disso, estudos futuros podem explorar variações mais sistemáticas de *prompts* e configurações de geração das LLMs, buscando reduzir a variabilidade inerente às saídas desses modelos.

Do ponto de vista metodológico, trabalhos futuros podem investigar a integração de *embeddings* gerados por LLMs com algoritmos clássicos de clusterização, como K-Means ou métodos hierárquicos. Nesse sentido, este trabalho já evidencia o potencial dessa integração ao demonstrar que representações mais ricas semanticamente contribuem para a formação de agrupamentos mais coerentes, indicando um caminho promissor para abordagens híbridas em tarefas de organização de conteúdo textual.

Referências

AGRAWAL, A.; FU, W.; MENZIES, T. What is wrong with topic modeling? and how to fix it using search-based software engineering. **Information and Software Technology**, v. 98, p. 74–88, 2018. ISSN 0950-5849. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584917300861>. Citado 3 vezes nas páginas 29, 35 e 41.

BASU, S. et al. Videopedia: Lecture video recommendation for educational blogs using topic modeling. In: . [S.l.: s.n.], 2016. ISBN 978-3-319-27670-0. Citado 2 vezes nas páginas 31 e 33.

BASU, S.; YU, Y.; ZIMMERMANN, R. Fuzzy clustering of lecture videos based on topic modeling. In: **2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)**. [S.l.: s.n.], 2016. p. 1–6. Citado 2 vezes nas páginas 30 e 33.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003. Disponível em: <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>. Citado na página 28.

BOUCHARD, L. **Mixture of Experts (MoE)**. 2024. <https://www.louisbouchard.ai/moe/>. Acessado em: 31 de janeiro de 2026. Citado 3 vezes nas páginas 7, 19 e 20.

CARON, M. et al. Deep clustering for unsupervised learning of visual features. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 132–149. Citado 2 vezes nas páginas 32 e 33.

DeepSeek-AI. Deepseek-moe: Towards efficient large language models with mixture of experts. **arXiv preprint arXiv:2401.06066**, 2024. Citado na página 27.

_____. **DeepSeek: Technical Overview**. 2025. <https://www.deepseek.com/blog>. Acessado em: 1 de fevereiro de 2026. Citado 2 vezes nas páginas 27 e 28.

DU, N. et al. Glam: Efficient scaling of language models with mixture-of-experts. **arXiv preprint**, arXiv:2112.06905, 2022. Disponível em: <https://arxiv.org/abs/2112.06905>. Citado na página 23.

FEDUS, W.; ZOPH, B.; SHAZEER, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. **arXiv preprint arXiv:2101.03961**, 2021. Disponível em: <https://arxiv.org/abs/2101.03961>. Citado na página 22.

GILLESPIE, T. Media technologies: essays on communication, materiality and society. In: **The Relevance of Algorithms**. [S.l.]: MIT Press, 2014. cap. 9, p. 167–193. Citado na página 14.

GOMES, P. C. T. **O que são LLMs (Large Language Models)?** 2024. <https://www.datageeks.com.br/llms/>. Acessado em: 07 de outubro de 2024. Citado na página 25.

Google DeepMind. Gemini: A family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023. Disponível em: <https://arxiv.org/abs/2312.11805>. Citado na página 27.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **Proceedings of the National Academy of Sciences**, v. 101, n. Suppl 1, p. 5228–5235, 2004. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.0306552101>. Citado na página 29.

HUANG, C.; HE, G. Text clustering as classification with large language models. **arXiv preprint arXiv:2410.00927**, 2024. Disponível em: <https://arxiv.org/abs/2410.00927>. Citado 2 vezes nas páginas 31 e 33.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, Springer, v. 2, n. 1, p. 193–218, 1985. Citado na página 37.

JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. **Bulletin de la Société Vaudoise des Sciences Naturelles**, 1901. Citado na página 40.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 31, n. 3, p. 264–323, set. 1999. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/331499.331504>. Citado na página 28.

JANSEN, A. et al. Large-scale audio event discovery in one million youtube videos. In: **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 786–790. Citado 2 vezes nas páginas 31 e 33.

LEVIATHAN, Y.; KALMAN, M.; MATIAS, Y. Fast inference from transformers via speculative decoding. **arXiv preprint arXiv:2211.17192**, 2022. Disponível em: <https://arxiv.org/abs/2211.17192>. Citado 2 vezes nas páginas 23 e 24.

LIORA. **All About Multi-Token Prediction**. 2025. <https://liora.io/en/all-about-multi-token-prediction>. Acessado em: 31 de janeiro de 2026. Citado 2 vezes nas páginas 8 e 24.

OLIVEIRA, N. R. de et al. Processamento de linguagem natural para identificação de notícias falsas em redes sociais: Ferramentas, tendências e desafios. In: **Minicursos do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais**. [S.l.]: Sociedade Brasileira de Computação - SBC, 2020. cap. 2, p. 51–100. Citado na página 17.

ONYEPUNUKA, U. et al. Multilingual analysis of youtube's recommendation system: Examining topic and emotion drift in the'cheng ho'narrative. In: **Text2Story@ ECIR**. [S.l.: s.n.], 2023. p. 25–35. Citado 2 vezes nas páginas 32 e 33.

OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023. Disponível em: <https://arxiv.org/abs/2303.08774>. Citado na página 26.

ORIOL, R. **Understanding LLMs: Mixture of Experts**. 2024. <https://dev.to/rogiia/understanding-llms-mixture-of-experts-jbm>. Acessado em: 31 de janeiro de 2026. Citado 2 vezes nas páginas 7 e 23.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. In: **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.: s.n.], 2007. Citado na página 39.

SANCHES, G. R. D. S. Trabalho de Conclusão de Curso Técnico, **ENTENDENDO O ALGORITMO DO YOUTUBE: ESTRATÉGIAS PARA CRIAÇÃO DE PRESENÇA ONLINE**. Londrina, Paraná: [s.n.], 2023. Citado na página 14.

scikit-learn developers. **Adjusted Mutual Information**. 2024. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html>. Acessado em: 28 de janeiro de 2026. Citado na página 38.

SHAZEER, N. et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. **arXiv preprint**, arXiv:1701.06538, 2017. Disponível em: <https://arxiv.org/abs/1701.06538>. Citado na página 22.

TURCU, G. et al. Video transcript indexing and retrieval procedure. In: **2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)**. [S.l.: s.n.], 2019. p. 1–6. Citado 2 vezes nas páginas 31 e 33.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado 5 vezes nas páginas 7, 18, 21, 25 e 26.