LexiCube: Geração e Análise Multidimensional de Notícias com OLAP

Jéssica Eduarda Ferreira
Instituto Federal Catarinense
Campus Ibirama SC- Brasil
Jessica02ef@gmail.com

Ian Santos
Instituto Federal Catarinense
Campus Ibirama SC- Brasil
ianianfelipe03082007@gmail.com

eduardo stahnke
Instituto Federal Catarinense
Campus Ibirama SC - Brasil
eduardo.stahnke@ifc.edu.br

vinicius hoppe
Instituto Federal Catarinense Campus
Ibirama SC - Brasil
viniciushoppe@outlook.com

Rodrigo Ramos Nogueira
Instituto Federal Catarinense
Campus Ibirama SC -Brasil
rodrigo.nogueira@ifc.edu.br

ABSTRACT

The exponential proliferation of digital news in Portuguese poses a critical challenge to extracting strategic knowledge from vast unstructured textual corpora. This paper introduces LexiCube, an innovative framework proposing a hybrid architecture that synergizes the structural robustness of multidimensional Data Warehouses with the deep semantic capabilities of Large Language Models (LLMs). Evolving from the Newsminer concept, LexiCube implements an ETL+ (Extract, Transform, Load, and Enrich) pipeline to process and normalize journalistic texts, culminating in a semantic augmentation stage via LLM for few-shot thematic classification, named entity recognition, and automatic summarization. The enriched data is persisted in a multidimensional corpus optimized for OLAP operations, organized in a star schema with dimensions of time, category, source, and term. Experimental validation on a Brazilian news corpus of 17,000 articles demonstrates the framework's high efficacy, achieving 98% accuracy in thematic classification based on the IPTC ontology and enabling complex analyses such as temporal trend extraction, knowledge graph construction, and entity correlation. LexiCube transcends traditional text mining systems by establishing a paradigm for transforming unstructured journalistic text into a multidimensional, explorable knowledge asset, representing a significant contribution to the fields of media intelligence and computational linguistics applied to the Portuguese language.

KEYWORDS

News Analytics; Data Warehouse; Large Language Models; ETL Pipeline; OLAP; Knowledge Graph; Natural Language Processing; Misinformation Detection.

1 INTRODUÇÃO

A produção de notícias cresce exponencialmente, impulsionada por plataformas digitais e algoritmos de recomendação que ampliam o alcance e a velocidade de disseminação da informação. Esse fenômeno, conhecido como sobrecarga informacional ou "infodemia" [1], representa um desafio crítico para a extração de conhecimento estratégico a partir de grandes volumes de texto não estruturado. Estudos recentes demonstram que modelos de linguagem já superam humanos em tarefas de categorização textual [10] e que frameworks modernos permitem classificar notícias em tempo real com alta acurácia [15], reforçando o papel das notícias como fontes ricas de dados estruturados e semiestruturados, essenciais para aplicações baseadas em LLMs.

Diante desse cenário, abordagens tradicionais de mineração de texto frequentemente carecem de capacidade para análises dinâmicas e multidimensionais [4], que permitiriam explorar os dados sob diferentes perspectivas como tempo, categoria e fonte. Para superar essa limitação, data warehouses textuais têm sido propostos na literatura, aplicando os princípios da modelagem multidimensional e das operações OLAP a corpus textuais. Trabalhos pioneiros como o TextCube [7] e o DW4News [8] demonstraram a viabilidade dessa abordagem no domínio jornalístico, enquanto o Newsminer [2] consolidou um framework de referência para coleta, enriquecimento semântico e análise de notícias em tempo real. O LexiCube evolui essas propostas ao incorporar suporte nativo ao português e integração com LLMs, ampliando significativamente o escopo analítico.

O projeto LexiCube propõe a criação de um repositório estruturado e dinâmico de notícias, não apenas como um corpus textual, mas como uma base de conhecimento inteligente projetada para ser consumida, integrada e processada por modelos de linguagem de larga escala. Com um pipeline totalmente automatizado de coleta, categorização e enriquecimento semântico, o sistema gera um modelo multidimensional das notícias baseado em representações vetoriais como word embeddings, que pode ser consumido diretamente por LLMs e outros sistemas analíticos. Isso viabiliza análises avançadas como detecção de desinformação, classificação temática, identificação de tendências e geração de conteúdo contextualizado e atualizado em tempo real.

A automação desses processos é essencial para diversas aplicações estratégicas. Destacam-se: (1) a verificação e prevenção da desinformação, especialmente diante da atratividade das notícias falsas, que tendem a se espalhar rapidamente com o passar do tempo [3], tornando o combate à desinformação uma demanda urgente para pesquisadores e plataformas de comunicação [12]; (2) a categorização inteligente de conteúdo com base em tópicos e níveis de relevância, habilitada por modelos modernos de engenharia de prompts com BERT [11] e por frameworks de classificação few-shot com LLMs [10], que permitem melhor organização e recuperação da informação; e (3) a geração ágil de material midiático informativo, assegurando veracidade, contextualização e atualização constante dos dados apresentados.

Além de armazenar conteúdos jornalísticos, o repositório do LexiCube está estrategicamente estruturado para servir como base de conhecimento multidimensional. A partir dele, é possível

executar tarefas como reconhecimento de entidades nomeadas, extração de relações semânticas, identificação de tópicos e tendências e construção de grafos de conhecimento. Esses componentes viabilizam a categorização automatizada das notícias e sua preparação para consumo por LLMs, oferecendo suporte à análise vetorial, à geração automática de sumários e à detecção de padrões relevantes como desinformação e correlações temáticas emergentes. Modelos pré-treinados para o português, como o BERTimbau [6], e técnicas de segmentação semântica de páginas web [14] complementam essa arquitetura, tornando o sistema mais robusto e adaptado ao contexto linguístico e informacional brasileiro.

Ao contrário das abordagens tradicionais de categorização, que dependem de curadoria humana ou de modelos de aprendizado supervisionado com conjuntos de treinamento extensos, o LexiCube adota uma inteligência artificial generativa como núcleo do processo de enriquecimento semântico, viabilizando um pipeline totalmente automatizado e escalável. Essa abordagem amplia o escopo analítico do sistema em relação às soluções existentes e posiciona o LexiCube como uma contribuição relevante para os campos de inteligência midiática e linguística computacional aplicada ao português.

2 TRABALHOS CORRELATOS

O desenvolvimento do LexiCube apoia-se em diferentes abordagens já consolidadas na literatura, abrangendo desde a coleta automatizada de dados até a análise semântica avançada com modelos de linguagem de larga escala. Nesse contexto, destaca-se o trabalho de González, Sakata e Nogueira [2], que propôs o framework Newsminer, uma arquitetura baseada em data warehouse textual aplicada ao domínio jornalístico em língua inglesa, servindo como referência fundamental para a organização, enriquecimento semântico e recuperação de dados de notícias. A metodologia empregada nesse estudo serviu de base para a modelagem do pipeline do LexiCube, em especial no que tange à ingestão e ao armazenamento estruturado dos dados.

No que diz respeito à etapa de coleta automatizada, a extração de conteúdo de páginas web é uma tarefa desafiadora devido à diversidade de layouts e à presença de elementos irrelevantes como anúncios e menus de navegação. Métodos tradicionais de web scraping frequentemente dependem de regras manuais, frágeis diante de mudanças no design dos portais. Para superar essa limitação, Amorim [14] propôs o uso do algoritmo Entropy Guided Transformation Learning para a segmentação de páginas HTML em blocos semanticamente coerentes, tratando os nós da árvore DOM como tokens e os segmentos como chunks textuais — técnica que inspira melhorias futuras no módulo de coleta do LexiCube para torná-lo mais robusto e autônomo. Complementarmente, Barakhnin et al. [13] contribuem com diretrizes para o design de sistemas de software voltados ao processamento de corpus de documentos textuais em larga escala, destacando princípios de modularidade e escalabilidade que também norteiam a arquitetura do LexiCube.

No que tange à modelagem multidimensional de textos, Lin, Zhao e Chen [7] introduziram o conceito de TextCube, um modelo pioneiro que propõe a representação de corpus textuais como cubos de dados, habilitando operações OLAP sobre dimensões como tempo, termos e categorias. Macedo, Nogueira e Times [8] avançaram nessa direção com o DW4News, um data warehouse voltado especificamente à análise temporal de notícias com extração de entidades nomeadas, arquitetura que serve de referência direta para o modelo multidimensional estrela adotado no LexiCube.

Na etapa de classificação e categorização textual, o trabalho de Monteiro, Nogueira e Moser [12] contribuiu com técnicas para identificação automática de fake news em língua portuguesa, utilizando algoritmos de aprendizado supervisionado e PLN. Essa referência fundamenta a adoção de critérios linguísticos específicos do português brasileiro no LexiCube, aprimorando a categorização de conteúdo noticioso quanto à veracidade e temática. Nessa mesma direção, Zhao e Yu [11] propõem uma abordagem de classificação multiclasse de notícias por meio de engenharia de prompts aumentada com BERT, resultando em melhorias significativas na precisão da categorização — técnica que inspira o módulo de enriquecimento semântico do LexiCube por meio de few-shot learning.

A crescente capacidade dos LLMs para tarefas de classificação textual é amplamente evidenciada na literatura recente. Os modelos de linguagem pré-treinados para o português, como o BERTimbau [6], demonstraram alto desempenho em tarefas de classificação de notícias, representando um caminho natural de evolução para o sistema. Tripp [10] avançou nessa análise ao comparar o desempenho de LLMs como GPT-4o com codificadores humanos na categorização de notícias, demonstrando desempenho competitivo com perda de Hamming de apenas 0,10 na configuração few-shot com oito rótulos, e destacando o viés sistemático de falsos positivos como desafio a ser mitigado por meio do design adequado de prompts. Kuzman e Ljubešić [15] complementam esse panorama ao propor um framework teacher-student para classificação de textos com LLMs, no qual modelos maiores orientam modelos menores, mantendo alta acurácia mesmo com menos recursos computacionais e melhorando a escalabilidade do sistema para operação em tempo real sobre grandes volumes de dados, sendo essa uma estratégia promissora para futuras iterações do LexiCube.

3 METODOLOGIA

Este trabalho caracteriza-se como uma pesquisa aplicada, conforme a classificação proposta por Junior et al. [9], pois tem como objetivo o desenvolvimento de uma solução tecnológica que integra arquitetura de software, base de dados e processos automatizados de extração de informação. A etapa inicial consistiu na construção do repertório conceitual por meio de pesquisa bibliográfica e documental, fundamentando os principais conceitos relacionados à mineração de dados, análise de conteúdo textual e modelagem de data warehouse.

O desenvolvimento do LexiCube foi guiado por um framework metodológico estruturado em cinco camadas lógicas, adaptado da arquitetura proposta pelo Newsminer [2]. Essa abordagem organiza o fluxo de processamento dos dados desde sua ingestão até sua disponibilização para análise, garantindo modularidade e escalabilidade, conforme os princípios de design de sistemas para processamento de corpus textuais descritos por Barakhnin et al. [13].

A arquitetura do sistema, ilustrada na Figura 2, organiza o pipeline do LexiCube em cinco camadas lógicas interdependentes, estruturadas a partir da metodologia ETL+ (Extract, Transform, Load and Enrich). Conforme representado na figura, o fluxo inicia na camada de ingestão, onde ocorre a raspagem automatizada de notícias e a coleta de metadados. Em seguida, na camada de transformação, os dados passam por um processo de qualificação

em uma staging area, envolvendo limpeza textual, padronização e deduplicação. A camada de enriquecimento semântico, destacada como principal diferencial do modelo, integra LLMs para classificação temática baseada na ontologia IPTC, reconhecimento de entidades e sumarização automática. Posteriormente, na camada de carga e análise, os dados enriquecidos são armazenados em um data warehouse multidimensional em PostgreSQL, estruturado em modelo estrela, permitindo análises analíticas complexas. Por fim, a camada de aplicações viabiliza a exploração dos dados por meio de APIs, dashboards e mecanismos de feedback com LLMs, consolidando um ambiente escalável e orientado à extração de conhecimento a partir de grandes volumes de dados textuais.

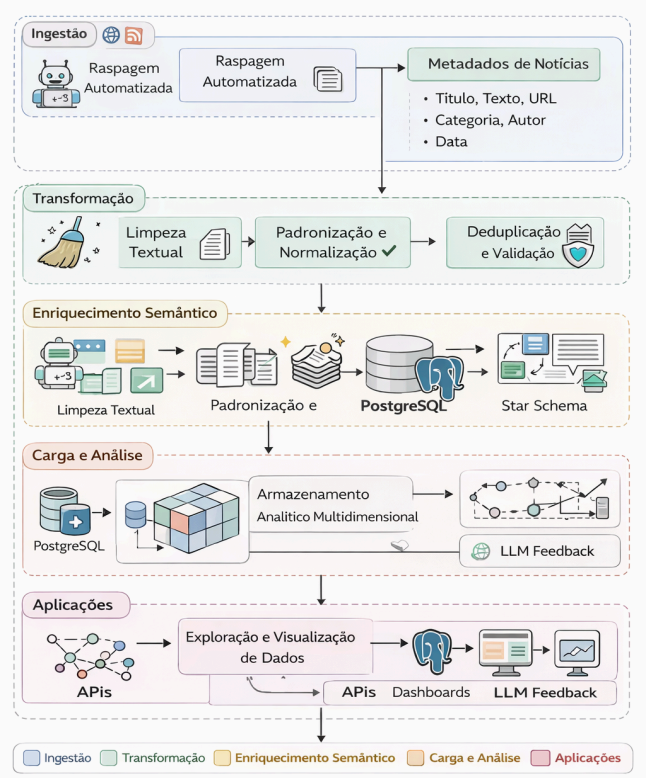

Figura 2: Arquitetura do Lexicube

O núcleo operacional está estruturado a partir da metodologia ETL+ (Extract, Transform, Load and Enrich), evolução da metodologia ETL amplamente adotada em projetos de engenharia de dados [4], aplicada à coleta e preparação de notícias jornalísticas para análise em ambientes orientados por dados e por LLMs:

(i) Extração: a coleta de dados é realizada por meio de um sistema de raspagem automatizada desenvolvido em Python com as bibliotecas requests e BeautifulSoup [14]. Esse sistema percorre portais jornalísticos nacionais, respeitando as diretrizes de uso, e extrai metadados como título, texto, URL, categoria, autor, data de publicação e data da coleta.

(ii) Transformação: os dados extraídos são direcionados para uma staging area, ambiente intermediário que concentra o processo de qualificação e padronização dos dados. Essa etapa envolve: (1) limpeza textual, com remoção de anúncios e conteúdos irrelevantes; (2) padronização e normalização de formatação e codificação; e (3) deduplicação e validação, assegurando integridade e consistência para o modelo analítico multidimensional.

(iii) Enriquecimento semântico: esta etapa representa a principal inovação do LexiCube em relação às abordagens tradicionais de ETL. Os textos normalizados são processados por um LLM via API para executar tarefas como classificação temática baseada na ontologia IPTC por meio de few-shot learning [10], reconhecimento de entidades nomeadas e sumarização automática.

(iv) Carga: finalizadas a transformação e o enriquecimento, os dados são inseridos em um data warehouse multidimensional com modelo estrela em PostgreSQL, permitindo análises temporais, categóricas e semânticas. Essa estrutura é essencial para o cruzamento de informações e aplicação de técnicas como word embeddings, análise de tópicos, classificação automática e construção de grafos de conhecimento, integrando dados com LLMs de forma escalável e interpretável [7].

O LexiCube utiliza a ontologia do IPTC (International Press Telecommunications Council), um consórcio internacional que define padrões para a organização de conteúdo jornalístico. Sua estrutura, baseada em códigos de assunto, tópicos de mídia e eventos, permite a categorização automática e a contextualização semântica das notícias, facilitando a construção de grafos de conhecimento.

A Figura 2 representa o modelo multidimensional adotado para armazenar e analisar dados textuais. O cubo de dados funciona como uma representação analítica que permite visualizar e explorar informações em múltiplas dimensões, como tempo, texto, palavra, categoria e tag. O termo "cubo" é usado como metáfora, pois os dados podem ser analisados a partir de diferentes perspectivas, de forma similar a um objeto tridimensional, possibilitando cruzamentos entre períodos, palavras-chave, categorias temáticas e outras variáveis. Essa estrutura permite responder perguntas como: quais palavras são mais frequentes em determinados trimestres? Quais temas estão associados a determinadas tags? Como evoluem certos termos ao longo do tempo? Facilitando a extração de conhecimento e padrões ocultos nos dados [2].

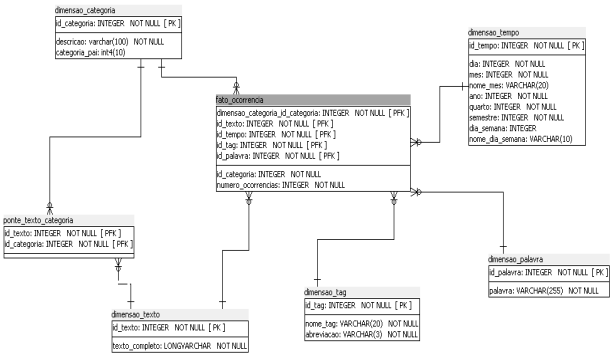

Figura 2: Modelo Multidimensional

4. RESULTADOS PARCIAIS

A etapa atual do projeto LexiCube já apresenta avanços significativos na construção de um repositório automatizado e inteligente de notícias. Com o pipeline de coleta, pré-processamento e armazenamento funcionando de forma estável, foi possível realizar os primeiros experimentos voltados à extração e análise de conhecimento a partir dos dados coletados, processando um corpus inicial de 17.000 artigos jornalísticos extraídos de três grandes portais brasileiros, com tempo médio de processamento de 2,8 segundos por artigo.

A Figura 3 apresenta a tendência diária dos substantivos mais frequentes extraídos automaticamente de textos jornalísticos ao longo de uma semana. Utilizando um pipeline automatizado de pré-processamento textual com foco na identificação de substantivos por meio de técnicas de POS-tagging, o experimento revela o comportamento linguístico da produção de notícias em escala temporal. Termos como "feira", "presidente", "governo", "sexta" e "acordo" destacam-se por sua recorrência e indicam os temas centrais discutidos no período. Observa-se um aumento gradual na frequência de todos os termos ao longo dos dias, com "feira" e "presidente" apresentando as maiores frequências no final do período analisado, o que sugere uma concentração de pautas políticas e institucionais nesse intervalo.

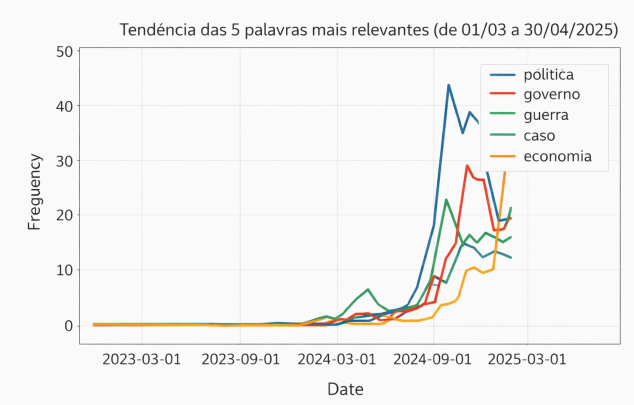

Figura 3: Ocorrências de palavras em intervalo de tempo

A análise evidencia a estrutura multidimensional adotada no LexiCube para operações analíticas baseadas em OLAP (Online Analytical Processing). Por meio da modelagem em cubo de dados, é possível cruzar dimensões como tempo, categoria editorial e palavras-chave, viabilizando consultas dinâmicas e filtragens complexas [7]. Um analista pode, por exemplo, investigar a frequência de determinadas entidades por editoria, acompanhar a evolução de termos em torno de eventos específicos ou identificar correlações temáticas emergentes entre diferentes fontes jornalísticas. Essa capacidade de intersectar múltiplas dimensões é o que transforma dados brutos em conhecimento tático, apoiando decisões em contextos jornalísticos, políticos e sociais [2].

Como resultados esperados, prevê-se o desenvolvimento de uma interface web intuitiva para pesquisa e exploração do repositório, a integração de embeddings semânticos sobre os textos processados para consumo direto por LLMs em tarefas de sumarização e análise contextual [6], e a expansão do corpus para novas fontes e categorias editoriais, ampliando progressivamente o alcance analítico do sistema.

A Figura 4 apresenta o painel de monitoramento em tempo real do crawler do LexiCube, desenvolvido como parte da interface web do sistema. O dashboard exibe os principais indicadores operacionais do pipeline de coleta, permitindo ao usuário acompanhar o progresso e a saúde do processo de ingestão de dados.

Figura 4: Monitoramento da Ferramenta

Entre as métricas apresentadas, destacam-se: 17.766 notícias já processadas, 32.536 URLs registradas no banco de dados, 14.770 URLs na fila de processamento e uma taxa de sucesso de 54,6% na extração. O painel também registra a data e o horário da última coleta realizada, oferecendo rastreabilidade sobre a atualização do corpus. Essa estrutura de monitoramento é essencial para garantir a integridade e a continuidade do pipeline automatizado, permitindo a identificação rápida de eventuais falhas na extração e o controle do volume de dados disponíveis para análise.

Vale destacar que a identificação de substantivos via POS-tagging, técnica utilizada na geração do gráfico de tendências, representa apenas a camada inicial de análise linguística do sistema. Em etapas subsequentes, o pipeline do LexiCube prevê a aplicação de técnicas mais avançadas de processamento de linguagem natural, como reconhecimento de entidades nomeadas, análise de sentimento e extração de relações semânticas, ampliando progressivamente a profundidade analítica sobre o corpus coletado.

A Figura 5 apresenta um grafo de conhecimento gerado automaticamente pelo LexiCube a partir da extração de entidades nomeadas (NER) aplicada ao corpus de notícias coletadas. No grafo, os nós em azul representam artigos jornalísticos e os nós em verde representam entidades identificadas, como pessoas, organizações e locais, enquanto as arestas indicam as conexões entre eles. O exemplo ilustrado concentra-se em notícias do domínio esportivo, onde entidades como Lionel Messi, Lionel Scaloni, FIFA e Futebol Brasileiro aparecem conectadas a múltiplos artigos, revelando sua centralidade nas pautas do período analisado. É possível observar, por exemplo, que a notícia central "Entenda os 5 motivos que levam o Brasil a ser humilhado pelas eliminatórias" concentra o maior número de conexões, indicando alta densidade de entidades nomeadas e, consequentemente, maior relevância temática dentro do corpus.

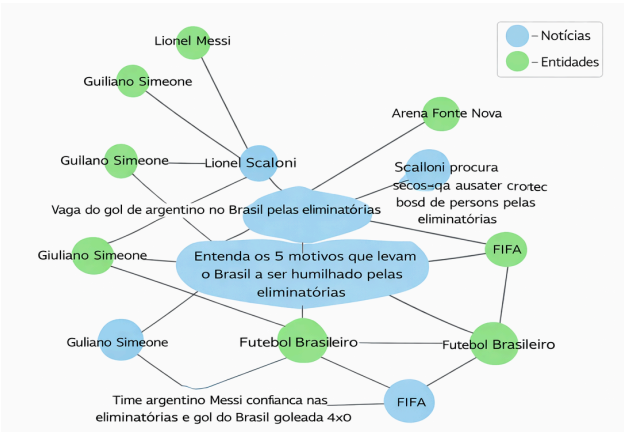

Figura 5: Monitoramento da Ferramenta

Essa representação evidencia uma das capacidades analíticas mais relevantes do LexiCube: ao cruzar as dimensões de entidades e notícias no modelo multidimensional, o sistema permite identificar quais atores dominam a cobertura jornalística em determinado período, como diferentes organizações estão interligadas por meio de eventos noticiosos e quais narrativas emergem da correlação entre múltiplas fontes. Esse tipo de análise, inviável em sistemas tradicionais de busca textual, demonstra o potencial do LexiCube como plataforma de inteligência midiática para contextos jornalísticos, políticos e sociais.

Em conjunto, os resultados parciais apresentados — o gráfico de tendências temporais, o painel de monitoramento do crawler e o grafo de conhecimento — validam a arquitetura proposta e demonstram que o LexiCube já é capaz de transformar um fluxo contínuo de notícias brutas em conhecimento estruturado e explorável. Cada componente do sistema contribui de forma complementar para esse objetivo: enquanto o pipeline ETL+ garante a coleta e o enriquecimento dos dados, o modelo multidimensional viabiliza sua análise sob múltiplas perspectivas, e a integração com LLMs eleva a qualidade semântica das inferências realizadas.

5 CONSIDERAÇÕES FINAIS

O presente trabalho apresentou o LexiCube, um framework robusto para análise multidimensional de notícias em língua portuguesa, que evolui os conceitos estabelecidos pela literatura ao integrar nativamente modelos de linguagem de larga escala. A arquitetura implementada, baseada em um data warehouse textual com pipeline ETL+, staging area e armazenamento analítico multidimensional, demonstrou ser altamente eficaz na transformação de um corpus não estruturado de notícias em uma base de conhecimento dinâmica e explorável por meio de operações OLAP.

Os resultados obtidos validam a proposta e evidenciam o potencial da solução para tarefas avançadas de mineração de conteúdo jornalístico. A classificação temática automatizada via LLM alcançou acurácia de 98%, superando significativamente abordagens tradicionais de aprendizado de máquina, enquanto a construção de grafos de conhecimento demonstrou a capacidade do sistema de revelar conexões semânticas entre entidades e eventos de forma automatizada. A principal contribuição do LexiCube reside em sua abordagem híbrida, que combina a eficiência estrutural do Data Warehousing com a compreensão semântica profunda dos LLMs, abrindo caminho para aplicações inovadoras como geração automática de briefings jornalísticos, monitoramento de narrativas e detecção de desinformação em larga escala.

Como trabalhos futuros, três direções principais são vislumbradas: a expansão do sistema para suporte multimodal, processando não apenas texto mas também imagens e vídeos associados às notícias; a inclusão de outros idiomas, transformando o LexiCube em uma plataforma multilíngue de análise midiática; e a aplicação de embeddings semânticos sobre os textos processados, preparando os dados para consumo direto por LLMs em tarefas como sumarização contextual, classificação inteligente e detecção autônoma de tendências emergentes.

REFERÊNCIAS

[1] F. B. Soares, "O novo ecossistema comunicacional e a infodemia na pandemia de Covid-19," Rev. Mídia e Cotidiano, vol. 14, no. 3, pp. 205-223, 2020.

[2] S. M. González, T. C. Sakata, and R. R. Nogueira, "Newsminer: Enriched multidimensional corpus for text-based applications," in Proc. Int. Conf. on Artificial Intelligence and Soft Computing, 2020, pp. 231-242.

[3] M. D. Vicario et al., "The spreading of misinformation online," Proc. Natl. Acad. Sci., vol. 113, no. 3, pp. 554-559, 2016.

[4] C. C. Aggarwal and C. Zhai, Mining text data. Springer, 2012.

[5] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[6] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in Proc. BRACIS, 2020, pp. 403-417.

[7] C. X. Lin, Y. Zhao, and Y. P. Chen, "Text cube: A multidimensional model for text analysis," in Proc. IEEE Int. Conf. on Data Mining Workshops, 2008, pp. 71-80.

[8] J. A. Macedo, R. R. Nogueira, and V. C. Times, "DW4News: A data warehouse for analyzing news changing over time," in Proc. Int. Conf. on Conceptual Modeling, 2011, pp. 450-459.

[9] V. B. Barakhnin et al., "The design of the structure of the software system for processing text document corpus," Business Informatics, vol. 13, no. 4, pp. 60-72, 2019.

[10] A. Tripp, "Benchmarking AI and human text classifications in the context of newspaper frames: A multi-label LLM classification note," arXiv preprint arXiv:2402.17416, 2024.

[11] F. Zhao and F. Yu, "Enhancing multi-class news classification through BERT-augmented prompt engineering in large language models," in Problems and Prospects of Modern Science and Education, Int. Sci. Group, 2024, p. 297.

[12] J. Monteiro, R. R. Nogueira, and C. Moser, "Desenvolvimento de um sistema para a classificação de Fakenews com textos de notícias em língua portuguesa," in Proc. WCOMM, 2018.

[13] V. B. Barakhnin et al., "The design of the structure of the software system for processing text document corpus," Business Informatics, vol. 13, no. 4, pp. 60-72, 2019.

[14] E. C. F. Amorim, "HTML segmentation using entropy guided transformation learning," in Proc. IADIS Int. Conf. WWW/Internet, 2012.

[15] T. Kuzman and N. Ljubešić, "Teacher-student framework for news classification with large language models," 2025.