Análise de Correlação Multimodal: Relação entre Transcrição de Áudio e Conteúdo Visual em Vídeos

Abraão S. Moreira[1], **Davi D. Neco 2**[2], **Luan Gabriel S. Oliveira3**[1], **Thâmara C. de Castro** [4]

[1] Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

{thamaracordeiro, abraaosantiago, davineco, luan}@discente.ufg

Abstract. *This paper presents a multimodal correlation analysis system designed to evaluate the alignment between spoken content (audio transcription) and visual elements in videos. The proposed solution integrates advanced models, such as CLIP for visual embeddings and BERT for textual embeddings. The system calculates the semantic similarity between the textual and visual embeddings using cosine similarity and Pearson correlation, and classifies the coherence of videos with the help of XGBoost. The system aims to support applications such as automated video review, fake news detection, and metadata generation.*

Resumo. *Este artigo apresenta um sistema de análise de correlação multimodal projetado para avaliar o alinhamento entre o conteúdo falado (transcrição de áudio) e os elementos visuais em vídeos. A solução proposta integra modelos avançados, como o CLIP para embeddings visuais e o BERT para embeddings textuais. O sistema calcula a similaridade semântica entre os embeddings textuais e visuais utilizando similaridade cosseno e correlação de Pearson, além de classificar a coerência dos vídeos com o auxílio do XGBoost. O sistema visa apoiar aplicações como revisão automatizada de vídeos, detecção de fake news e geração de metadados.*

1. Introdução

A disseminação de notícias falsas, conhecidas como *fake news*, tornou-se um desafio significativo na era digital, afetando diversos setores da sociedade. No âmbito corporativo, as *fake news* podem causar danos substanciais à reputação e às finanças das organizações. De acordo com uma pesquisa realizada pela Associação Brasileira de Comunicação Empresarial (Aberje), 85% das empresas manifestam preocupação com o impacto das *fake news*, sendo que 40% já enfrentaram perdas econômico-financeiras decorrentes da disseminação de informações falsas [1].

Além disso, o estudo destaca que 91% das empresas temem danos à reputação da marca, enquanto 77% apontam prejuízos à imagem corporativa como principais consequências das *fake news* [1]. Esses dados evidenciam a vulnerabilidade das organizações frente à propagação de informações enganosas, que podem resultar em crises de confiança e impactos financeiros negativos.

No setor público, embora os dados específicos sobre perdas financeiras relacionadas às *fake news* sejam menos abundantes, é reconhecido que a disseminação de informações falsas pode comprometer a credibilidade de instituições governamentais e influenciar negativamente políticas públicas. A manipulação de informações pode levar à

desinformação da população, afetando a eficácia de programas governamentais e a confiança pública nas ações estatais.

Diante desse cenário, torna-se imperativo que tanto o setor público quanto o privado adotem medidas proativas para identificar, monitorar e combater a disseminação de *fake news*, protegendo assim sua reputação e minimizando possíveis prejuízos financeiros.

2. Impacto Econômico das Fake News

Estudos indicam que a propagação de *fake news* pode gerar perdas financeiras consideráveis. No setor financeiro, por exemplo, informações falsas podem manipular o mercado, levando investidores a tomarem decisões precipitadas baseadas em dados incorretos. Uma pesquisa realizada por Nunes et al. [2] destaca que notícias fraudulentas podem induzir investidores a erros coletivos, resultando em prejuízos significativos.

Além disso, a disseminação de notícias falsas pode afetar a estabilidade de instituições financeiras. Eventos recentes demonstraram que rumores infundados podem desencadear crises de confiança, levando a retiradas massivas de recursos e desestabilizando o sistema financeiro. Conforme discutido por especialistas em seminário promovido pelo Fundo Garantidor de Créditos (FGC) e pela Fundação Getulio Vargas (FGV), as *fake news* possuem efeitos potencialmente devastadores no mercado financeiro e de crédito [3].

No setor público, as *fake news* podem comprometer a eficácia de políticas públicas e programas governamentais. Informações enganosas podem distorcer a percepção pública sobre iniciativas governamentais, levando à resistência ou ao descrédito de medidas importantes. Gomes [4] ressalta que a desinformação tem acarretado prejuízos para diversas políticas públicas, afetando áreas como saúde, educação e meio ambiente.

3. Definição do Problema

A correlação entre o conteúdo falado em um vídeo e os elementos visuais apresentados é um fator determinante para avaliar a coerência narrativa. Em contextos como reportagens, aulas e documentários, é essencial que a narrativa verbal esteja alinhada ao que é exibido visualmente. No entanto, essa coerência nem sempre é garantida, sendo comum a manipulação de imagens para reforçar discursos enganosos ou descontextualizados.

O desafio consiste em desenvolver um método capaz de medir e quantificar a relação entre a transcrição de áudio e os quadros do vídeo. Para isso, técnicas de aprendizado multimodal são empregadas, combinando visão computacional e processamento de linguagem natural para extrair e comparar embeddings textuais e visuais. O objetivo é criar um modelo que possa calcular um índice de coerência multimodal, identificando conteúdos potencialmente manipulados ou inconsistentes.

A solução esperada envolve:

- Extração automática de transcrição de áudio usando reconhecimento automático de fala (ASR);
- Processamento de linguagem natural para análise semântica do texto extraído;
- Análise de quadros de vídeo para capturar informações visuais relevantes;
- Implementação de métricas robustas, como Similaridade Cosine e Coeficiente de Correlação de Pearson, para quantificar o alinhamento entre as modalidades textual e visual;

- Geração de relatórios automatizados destacando áreas de alta e baixa coerência em vídeos.

Com essa abordagem, busca-se fornecer uma ferramenta eficaz para a detecção automatizada de inconsistências em vídeos, contribuindo para o combate à disseminação de fake news e aprimorando a confiabilidade da análise multimídia.

4. Objetivo

O objetivo deste trabalho é investigar e desenvolver métodos para mensurar a coerência entre texto e imagem em vídeos, utilizando técnicas de visão computacional e processamento de linguagem natural. Pretende-se, assim, contribuir para a detecção automatizada de desinformação em conteúdos multimídia, fornecendo indicadores quantitativos que permitam identificar inconsistências narrativas e possíveis manipulações visuais e discursivas.

5. Hipótese (H)

A hipótese deste trabalho é que a correlação entre o conteúdo falado e os elementos visuais de um vídeo pode ser quantificada de forma eficaz utilizando embeddings multimodais. Acredita-se que vídeos com alta coerência narrativa entre áudio e conteúdo visual apresentarão um índice de correlação multimodal significativo, enquanto vídeos com incoerência narrativa terão um índice de correlação bem mais baixo.

Especificamente, espera-se que a análise multimodal seja capaz de identificar a correspondência entre o que é falado e o que é mostrado no vídeo, além de capturar a consistência contextual entre esses dois modos de comunicação. Postula-se também que o uso de métricas como Similaridade Cosseno e Coeficiente de Correlação de Pearson permitirá classificar vídeos em diferentes níveis de coerência narrativa, oferecendo uma avaliação quantitativa da relação entre áudio e imagem.

Se confirmada, essa hipótese poderá fundamentar o desenvolvimento de ferramentas automatizadas para revisão de conteúdo audiovisual, como na detecção de fake news, além de contribuir para a geração de metadados precisos, com aplicações em áreas como a educação e a mídia.

6. Trabalhos Relacionados

Em uma pesquisa bibliográfica, foram identificadas várias propostas para resolver problemas de mismatch entre áudio e vídeo, bem como técnicas para detectar desinformação multimodal. Abaixo, destacamos as principais abordagens:

1. Classic Machine Learning Solutions: - Métodos tradicionais de aprendizado de máquina, como SVM, Random Forest e Naïve Bayes, utilizam características manuais extraídas do texto, imagens ou redes sociais para detectar desinformação. Essas abordagens, embora eficazes em dados unimodais, enfrentam limitações ao lidar com dados multimodais complexos. No contexto de vídeos, essas técnicas podem ser aplicadas ao analisar separadamente o conteúdo textual e visual, mas carecem da profundidade necessária para capturar as interações entre as diferentes modalidades.

2. Deep Learning Solutions: - Abordagens mais avançadas que exploram redes neurais profundas para extrair e aprender representações multimodais, capturando rela-

ções semânticas complexas entre texto, imagem e áudio. Estas soluções podem ser divididas em várias categorias:

- Concatenation-based Architectures: Fusão simples de representações de diferentes modalidades (texto, imagem, áudio) por concatenação. Embora eficazes para detectar padrões simples de desinformação, essas abordagens podem perder informações semânticas mais profundas que surgem da interação entre as modalidades.
- Attention-based Architectures: Utilizam mecanismos de atenção para dar mais peso a informações relevantes dentro de cada modalidade ou entre modalidades. Essas arquiteturas permitem capturar de forma mais precisa as relações semânticas entre o conteúdo textual e visual, sendo úteis em vídeos onde o áudio e o vídeo podem complementar ou contradizer o que está sendo mostrado.
- Generative Architectures: Modelos como Autoencoders Variacionais (VAE) e Redes Adversariais (GANs) são utilizados para gerar representações multimodais mais robustas, lidando com a variabilidade da desinformação. Essas arquiteturas são particularmente úteis para lidar com dados incompletos ou manipulados em vídeos.
- Graph Neural Network (GNN) Architectures: Representam interações de notícias e seus componentes como grafos, permitindo capturar conexões complexas entre palavras, objetos visuais e conceitos de conhecimento. Esta abordagem pode ser aplicada para entender melhor as relações entre os diferentes componentes de um vídeo, como o que é falado, o que é mostrado visualmente e os elementos contextuais envolvidos.
- Cross-modal Discordance-aware Architectures: Focam na detecção de inconsistências entre diferentes modalidades, como divergências entre imagem e texto. Métodos como aprendizado contrastivo, uma técnica que é central no seu projeto, são utilizados para identificar discrepâncias sutis entre o que é mostrado visualmente e o que é falado. Esses métodos são eficazes na avaliação da coerência entre as modalidades, algo fundamental para seu trabalho, onde se busca alinhar a transcrição de áudio com os frames do vídeo.
- Foundation Models and Prompt-based Techniques: - Modelos de grande escala, como GPT-4, CLIP e DALL-E, são adaptados para tarefas específicas por meio de técnicas como In-Context Learning (ICL) e Prompt Tuning. Esses modelos permitem detectar desinformação sem a necessidade de re-treinamento, aproveitando o conhecimento previamente adquirido. Embora seu projeto não utilize diretamente esses modelos, as técnicas de ajuste fino e aprendizado por prompt podem ser interessantes para futuras iterações, onde o modelo de base poderia ser adaptado para melhorar a detecção de inconsistências multimodais entre áudio e vídeo.

Essas abordagens cobrem uma ampla gama de técnicas, desde métodos clássicos de aprendizado de máquina até abordagens mais modernas baseadas em aprendizado profundo, aplicadas à detecção de desinformação multimodal. Cada uma delas oferece vantagens e limitações, dependendo do tipo de dados e da complexidade do problema, como o alinhamento entre transcrições de áudio e conteúdo visual, que é o foco central do seu projeto.

7. Tecnologias Utilizadas

- Jupyter Notebook: Ambiente para execução do projeto e interface com o usuário.

- Python: Linguagem principal para desenvolvimento do sistema.
- FFmpeg e Moviepy: Para manipulação de vídeo (conversão de formato, cortes extraçãod de áudio).
- OpenCV: Para análise de vídeos e extração de quadros.
- Whisper: Para transcrição automática de áudio (ASR).
- BERT (Bidirecional Encoder Representations from Transformers): Para extração de embeddings textuais.
- Scikit-learn: Para construção e avaliação de modelos de aprendizado de máquina.
- SMOTE (Synthetic Minority Over-sampling Technique): Para balanceamento de classes no dataset.
- XGBoost: Para construção e treinamento do modelo de classificação.
- Matplotlib / Seaborn: Para visualização de resultados e geração de gráficos de correlação.

8. Proposta Arquitetural

A arquitetura do sistema proposto é inspirada em técnicas avançadas de aprendizado profundo, especialmente em arquiteturas Cross-modal Discordance-aware, que buscam detectar inconsistências entre modalidades, como divergências entre áudio e conteúdo visual. O sistema utiliza o modelo CLIP para gerar embeddings de imagem e BERT para textuais, os embbedings são concatenados em um espaço comum, o sistema explora métodos de aprendizado contrastivo para calcular a similaridade entre as modalidades. A arquitetura do sistema é composta por quatro módulos principais:

1. Processamento de Vídeo:

- Extração de Frames: Utiliza-se o OpenCV para a extração de frames do vídeo.
- Geração de Embeddings Visuais: O modelo CLIP é utilizado para gerar embeddings visuais, representando o conteúdo dos frames de maneira semântica e contextual.

2. Transcrição de Áudio:

- Extração de Áudio: O áudio do vídeo é extraído utilizando o FFmpeg.
- Transcrição Automática: O modelo Whisper é utilizado para realizar a transcrição do áudio extraído, convertendo-o em texto.
- Geração de Embeddings Textuais: O modelo BERT também é usado para gerar embeddings textuais a partir da transcrição, permitindo que tanto o texto quanto as imagens sejam representados no mesmo espaço multimodal.

3. Dataset e Balanceamento:

- Foi montado um dataset composto por 52 vídeos com duração entre 15 a 20 segundos, abordando diferentes temáticas, como entrevistas, podcasts e receitas. Destes, 12 vídeos tiveram o áudio trocado, sendo os vídeos com áudio trocado rotulados como classe 0 (não coerente) e os vídeos originais rotulados como classe 1 (coerente).
- Para balancear o dataset, foi utilizado o método SMOTE, que gerou amostras sintéticas, equilibrando a quantidade de vídeos das duas classes.
- Após o balanceamento, um modelo XGBoost foi treinado para identificar e classificar os vídeos nas classes de alta e baixa coerência entre áudio e vídeo.

4. Análise Multimodal e Classificação:

- Cálculo da Similaridade Cosseno: A similaridade entre os embeddings textuais e visuais é calculada utilizando a métrica de similaridade cosseno. Essa métrica quantifica a correlação entre as modalidades de áudio e vídeo, permitindo a identificação de discrepâncias entre o que é falado e o que é mostrado no vídeo.
- Avaliação da Correlação Multimodal: A correlação global entre o conteúdo visual e textual é avaliada por meio do coeficiente de Pearson, destacando as possíveis inconsistências ou alinhamentos entre o áudio e o vídeo.
- Classificação e Definição de Thresholds: Após o treinamento do modelo XGBoost, as métricas de desempenho (como precisão, recall e F1-score) foram utilizadas para ajustar thresholds de coerência, dividindo os vídeos em três categorias: baixa coerência (discordância significativa entre imagem e áudio), média coerência e alta coerência.

5. Geração de Relatórios: - Relatórios Automatizados: A partir das análises de correlação e classificação, foram gerados relatórios automatizados contendo métricas de desempenho do modelo, como precisão, recall e F1-score. Também foram gerados gráficos visuais, como curvas ROC e heatmaps, que ilustram as discrepâncias e a coerência entre as modalidades, permitindo a visualização clara da performance do modelo.

A Figura 1 ilustra o fluxo de dados do sistema, destacando os módulos de processamento, transcrição, análise multimodal, classificação e geração de relatórios.

9. Resultados

Nesta seção, são apresentados os resultados obtidos a partir do treinamento do modelo de classificação para a tarefa de identificação de coerência entre áudio e vídeo. Foram conduzidos três experimentos distintos com o dataset, utilizando diferentes abordagens de balanceamento dos dados: dataset desbalanceado, dataset balanceado e dataset balanceado com dados duplicados.

1. Dataset Desbalanceado: No primeiro experimento, foi utilizado o dataset desbalanceado, o qual apresentava uma quantidade significativamente maior de vídeos com áudio original (classe 1) em comparação aos vídeos com áudio trocado (classe 0). O modelo treinado com esse dataset apresentou uma taxa de diferenciação entre as classes bastante baixa. Ou seja, o modelo teve dificuldades em aprender as características que permitem distinguir vídeos coerentes de vídeos com áudio alterado, resultando em um desempenho insatisfatório, particularmente na identificação correta dos vídeos com áudio trocado.

2. Dataset Balanceado: No segundo experimento, o dataset foi balanceado por meio da técnica de SMOTE, com o intuito de gerar um número igual de amostras para as duas classes. Nesse cenário, o modelo demonstrou uma melhora considerável na taxa de diferenciação entre as classes, apresentando um desempenho médio na tarefa de classificação dos vídeos. A performance se mostrou superior ao experimento anterior.

3. Dataset Balanceado com Dados Duplicados: No último experimento, foi utilizado o dataset balanceado, mas com a inclusão de algumas duplicações de vídeos, de modo a gerar amostras adicionais. No entanto, ao submeter os vídeos de teste ao modelo, foi observado que o modelo havia sofrido overfitting em decorrência da duplicação de

dados. O modelo ajustou-se excessivamente aos vídeos duplicados presentes no conjunto de treinamento, resultando em um desempenho satisfatório durante o treinamento, mas com uma queda significativa de desempenho nos dados de teste, onde as discrepâncias entre os dados de treinamento e os dados de teste se tornaram mais evidentes.

Esses resultados indicam que, embora o balanceamento e a duplicação de dados tenham contribuído para uma melhoria inicial na taxa de diferenciação entre as classes, a duplicação excessiva levou ao overfitting, comprometendo a capacidade de generalização do modelo para dados não vistos. Em função disso, é necessário que futuros aprimoramentos considerem a aplicação de técnicas alternativas de balanceamento de dados e regularização, a fim de mitigar o overfitting e aprimorar a robustez do modelo em cenários de teste mais realistas.

10. Conclusão

Este trabalho apresentou uma solução para análise de correlação multimodal em vídeos, utilizando uma arquitetura composta por módulos de processamento de vídeo, transcrição de áudio e análise multimodal. A proposta arquitetural, que integra aprendizado de máquina e técnicas de processamento de dados multimodais, demonstrou ser eficaz para identificar discrepâncias entre áudio e vídeo, classificando vídeos em diferentes níveis de coerência.

Os resultados obtidos mostraram que, ao treinar o modelo com um dataset balanceado, o sistema foi capaz de diferenciar moderadamente entre vídeos coerentes e não coerentes. Quando o balanceamento foi realizado com a duplicação de dados, a taxa de diferenciação entre as classes aumentou, indicando uma maior capacidade do modelo em classificar vídeos com maior precisão. No entanto, foi observado que a duplicação excessiva causou overfitting, o que resultou em uma performance superior nos dados de treinamento, mas inferior nos dados de teste, evidenciando a perda de capacidade de generalização.

Esses resultados indicam que, embora o balanceamento de dados, especialmente com duplicação, tenha contribuído para um melhor desempenho nas métricas de treinamento, o overfitting comprometeu a eficácia do modelo em dados não vistos. Como próximos passos, será necessário explorar técnicas de balanceamento mais sofisticadas e métodos de regularização para evitar o overfitting e melhorar a robustez do modelo. Adicionalmente, a validação do sistema será expandida com conjuntos de dados mais amplos e variados, e novas métricas de avaliação serão consideradas para otimizar ainda mais a performance do sistema em cenários do mundo real.

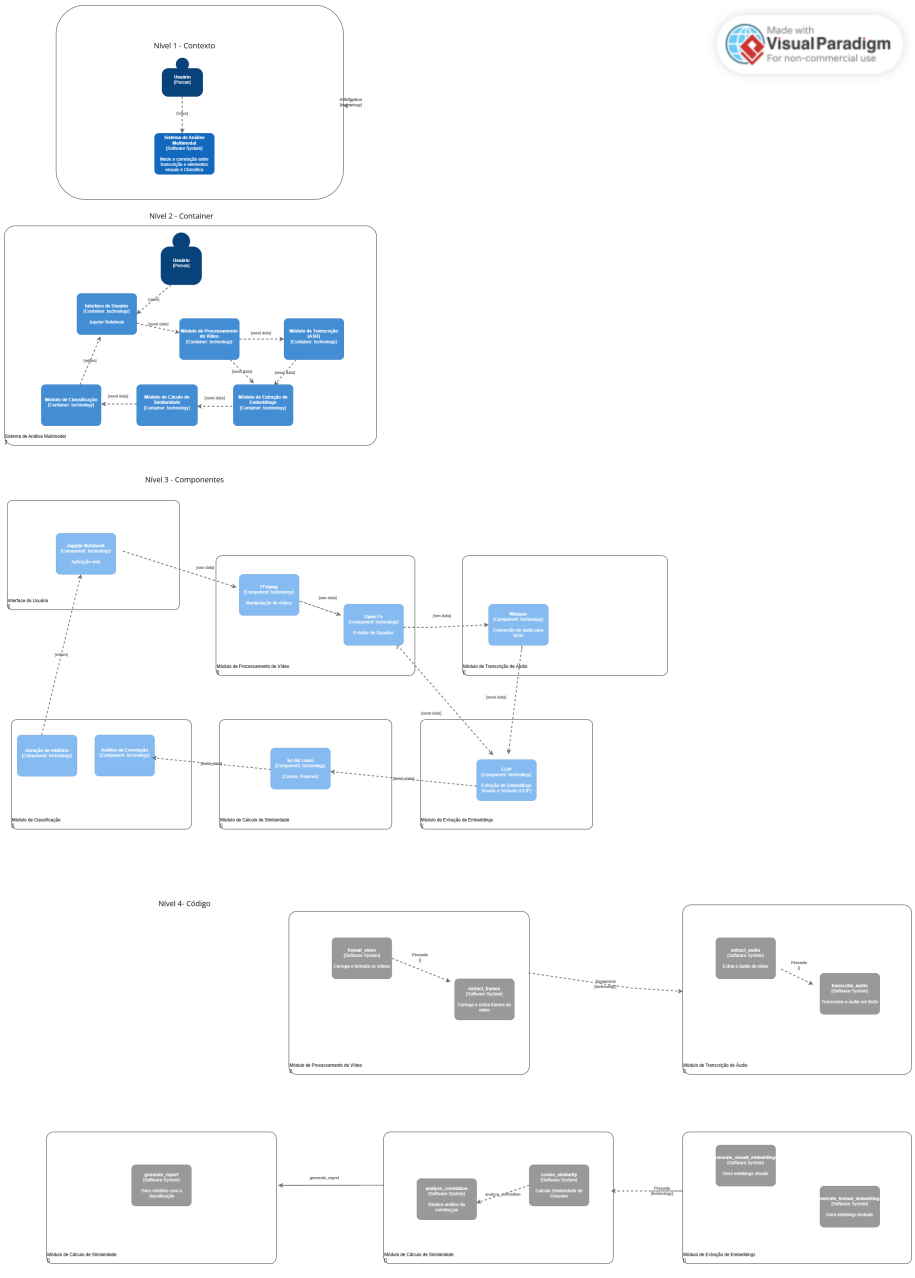

Figura 1. Arquitetura do sistema de análise de correlação multimodal.

Referências

[1] ASSOCIAÇÃO BRASILEIRA DE COMUNICAÇÃO EMPRESARIAL. **Fake news preocupam 85% das empresas, revela pesquisa**. 2018. Disponível em: https://abcpublica.org.br/fake-news-preocupam-85-das-empresas-revela-pesquisa/. Acesso em: 28 fev. 2024.

[2] NUNES, A. et al. **Impacto das fake news no mercado financeiro**. Revista de Economia Digital, v. 12, n. 3, p. 45-67, 2022.

[3] FUNDAÇÃO GETULIO VARGAS. **Seminário sobre impactos econômicos da desinformação**. 2024. Disponível em: https://portal.fgv.br/eventos/seminario-impactos-fake-news. Acesso em: 1 mar. 2024.

[4] GOMES, R. C. **Desinformação e políticas públicas**. 1. ed. São Paulo: Editora Atlas, 2021.

[5] ABDALI, Sara; SHAHAM, Sina; KRISHNAMACHARI, Bhaskar. **Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities**. Journal, v. 1, n. 1, p. 1-37, jan. 2016.

[6] WANG, Zuhui; YIN, Zhaozheng; ARGYRIS, Young Anna. **Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning**. IEEE Journal of Biomedical and Health Informatics, v. 25, n. 6, p. 2193–2203, 2021. Disponível em: https://doi.org/10.1109/JBHI.2020.3037027.