Estudo e modelagem para legendas de filmes infantis

Leoman Cássio Almeida dos Santos

Universidade do Estado da Bahia

Abstract. Este estudo apresenta uma análise computacional em larga escala de um corpus de legendas de filmes infantis, abrangendo 4 décadas. Utilizando técnicas de Processamento de Linguagem Natural (PLN), o trabalho foca em três eixos principais: (1) a transmissão de valores sociais ao decorrer das décadas; (2) a evolução linguística de termos e (3) a representação e variação das emoções nos diálogos. A metodologia baseia-se em pré-processamento textual, aplicação de dicionários lexicais e análise quantitativa de frequências, complementada por visualizações temporais.Os resultados apontam para a década de 2000 como um período de pico na frequência de termos associados a emoções, valores sociais como "Família"e "Amizade", e elementos de fantasia, seguido por uma normalização na década de 2010. Notavelmente, a proporção entre as emoções permaneceu estável ao longo do tempo, sugerindo que os filmes se tornaram mais expressivos em volume, mas não em sua estrutura emocional.

Keywords: Processamento de Linguagem Natural; Legendas; Processamento computacional

1. Introdução

O cinema infantil ultrapassa o mero entretenimento para se estabelecer como um dos mais poderosos agentes de socialização da cultura contemporânea. Dada sua atração e influência sobre as crianças, o cinema é um objeto de estudo fundamental para compreender as mensagens transmitidas à infância (BORGES; REIS, 2023).

Apesar disso, grande parte das análises sobre o tema se concentra em estudos de caso qualitativos, que, embora aprofundados, não conseguem capturar as tendências gerais que emergem ao longo de décadas. Diante desse cenário, o Processamento de Linguagem Natural (PLN) surge como a ferramenta ideal, permitindo a análise sistemática de grandes volumes de texto, como as legendas de filmes — um material que, por sua proximidade com a linguagem oral, é um excelente recurso para estudos linguísticos.(ESTIVALET et al., 2019)

Este trabalho aplica exatamente essa abordagem em um corpus inédito de 433 legendas (1980-2016) para contribuir no entendimento das transformações do cinema infantil. Nossa análise investiga três aspectos centrais: a evolução dos valores sociais, o mapeamento do vocabulário infantil e a variação na intensidade emocional dos diálogos.

2. Metodologia

Este estudo adotou uma abordagem quantitativa baseada em Processamento de Linguagem Natural (PLN) para analisar sistematicamente um corpus de legendas de filmes infantis. A metodologia seguiu um pipeline rigoroso em cinco etapas: aquisição do corpus, pré-processamento inicial, conversão e limpeza textual, agrupamento temporal e análise computacional.

2.1 Coleta e Construção do Corpus

O corpus foi constituído mediante um processo rigoroso de coleta e preparação de dados textuais. Inicialmente, realizaram-se o download e a descompactação de 433 arquivos de legendas em formato ZIP da plataforma Opensubtitles. Procedeu-se à conversão integral para o formato SubRip (SRT), com padronização de nomenclatura incluindo título do filme e ano de produção. Na sequência, executou-se a limpeza e extração do conteúdo textual, removendo metadados, numerações de cena e marcações temporais, preservando exclusivamente o conteúdo dialógico. Por fim, consolidaram-se os textos em quatro documentos organizados por decênios, formando a base unificada para análise.

A conversão das legendas para texto puro foi um passo essencial no pré-processamento. Realizou-se a limpeza integral dos metadados presentes nos arquivos de legenda, eliminando-se numerações de cena, marcações temporais e elementos de formatação, com preservação exclusiva do conteúdo dialógico original. Implementou-se ainda tratamento adequado para garantir a integridade dos caracteres especiais da língua portuguesa. Após esse processo de purificação textual, procedeu-se à consolidação das legendas por período cronológico, gerando quatro documentos textuais correspondentes às décadas estudadas, que constituíram a base unificada para as análises computacionais subsequentes.

2.2 Valores Sociais

A primeira análise focou na identificação e quantificação de valores sociais presentes no corpus. Implementou-se um sistema de categorização baseado em cinco dimensões conceptuais fundamentais, operacionalizadas através de

um léxico especializado. O protocolo analítico incorporou a segmentação do corpus em unidades contextuais, permitindo a captura de padrões de recorrência temática. A metodologia empregou contagem contextualizada de unidades lexicais predefinidas, habilitando a mensuração sistemática de construtos sociais complexos ao longo do eixo temporal investigado.

2.3 Dimensões Emocionais

A segunda análise dedicou-se ao mapeamento de dimensões emocionais através de uma estrutura baseada em categorias afetivas fundamentais. O protocolo utilizou 46 descritores lexicais validados para identificação de estados emocionais. A metodologia estendeu-se além da contagem lexical simples, incorporando análise de distribuição contextual e densidade emocional. Este enfoque permitiu capturar não apenas a frequência de marcadores emocionais, mas também seus padrões de ocorrência e variação diacrônica no tecido discursivo.

2.4 Evolução do Vocabulário

A terceira análise concentrou-se na evolução do vocabulário especializado mediante o monitoramento de 15 unidades lexicais nucleares. A metodologia empregou protocolos de correspondência padronizada com expressões regulares para garantir precisão na identificação de ocorrências semanticamente válidas. O enfoque analítico permitiu a quantificação de padrões de distribuição temporal e trajetórias de uso, assegurando a detecção confiável de transformações no vocabulário especializado ao longo do período estudado.

3. Resultados e discussão

4. Conclusões

Referências

BORGES, T. d. S.; REIS, M. S. A. Filmes infantis: parametro de analise a relacao de genero. *Computers, Materials and Continua*, v. 1, n. 1, 2023.

ESTIVALET, G. et al. Lexporbr infantil: uma base lexical tripartida e com interface web de textos ouvidos, produzidos, e lidos por crianças. In: *Symposium in Information and Human Language Technology - STIL*. [S.l.]: SBC, 2019.