Text Analysis of The Simpsons & Family Guy

Comparative Text Mining, Topic Modeling and Sentiment Analysis of Two Animated Sitcoms

Caicchiolo Giorgia, Carollo Daniele, Paganelli Gabriele, Pati Riccardo, Tossici Giorgia • University of Padova

1 216	36 + 23	9	3	1989–2024
Episodes	Seasons (S + G)	LDA Topics	STM Models	Period Covered

Overview

We conduct a comparative text analysis of **The Simpsons** (36 seasons, 784 episodes) and **Family Guy** (23 seasons, 432 episodes), two landmark American animated sitcoms. Script data was collected via web scraping from *Springfield!Springfield!*, then cleaned, lemmatized and preprocessed. The analysis pipeline covers **text mining** (frequency analysis, n-grams, keyness), **topic modeling** (LDA and STM) and **sentiment analysis** (AFINN lexicon-based). An episode-level crosswalk with IMDb ratings and synopses enriches interpretation throughout.

Data & Preprocessing

Sources: script text scraped from Springfield!Springfield!; IMDb ratings and episode summaries collected separately.

Preprocessing: tokenization, stopword removal, lemmatization (`textstem`), custom stoplist (low-frequency tokens, interjections, character names). Two corpora maintained per show: with and without character names.

Representations: per-episode Document-Term Matrices (DTM), sliding-window DTMs (80-word windows, step 40) for co-occurrence analysis, and `quanteda` DFMs for keyness.

Key lexical finding: *keyness* analysis (chi-squared) confirms that Family Guy uses markedly more provocative language — "hell", "damn", "Quahog" — while The Simpsons favours more neutral, family-friendly vocabulary. Top shared lemmas: *dad*, *love*, *kid*, *day*, *wait*.

Topic Modeling

LDA (via `topicmodels`) was fitted on three corpora: combined, Simpsons-only, Family Guy-only. Optimal k selected by balancing diagnostic metrics with qualitative coherence.

Corpus	k	Representative topics
Combined	9	Family, Christmas, School, News
Simpsons	7	Faith & Death, Homer & Work
Family Guy	5	Violence, Peter & Lois, Media

IMDb synopses validated topic assignments (e.g. the "Violence" topic's Christmas episode traces to Stewie attacking Santa Claus). Topics with character names reveal show-specific dynamics: Homer dominates The Simpsons across work, leisure and family; Stewie's prominence grows in later Family Guy seasons.

STM (`stm` package) extended the analysis with prevalence covariates — production year (continuous, factor, spline), US president, and showrunner — using *series* as content covariate. Key findings: year effects are nonlinear; the "Romantic Relationships" topic rose sharply post-2000; "Christmas Episodes" peaked in the late 1990s; showrunner effects are modest but Palladino (Family Guy) and Reiss (Simpsons) stand out.

Sentiment Analysis

Lexicon-based scoring (AFINN, -5 to $+5$) computed per episode and per season. Combined with topic assignments and IMDb ratings for trivariate exploration.

Series-level contrast: The Simpsons maintains a mildly positive mean sentiment throughout, with highest values in early seasons. Family Guy oscillates more widely and trends negative, especially in central seasons (4–12), recovering only in recent years.

Topic-level: "Family" and "Romantic Relationships" carry the highest median sentiment in both shows. "News/Current Events" and "Work & Money" skew negative. The Family Guy "Violence" topic shows the most extreme negative values.

Longitudinal: comparing equivalent topics across shows (Family, Couple, Leisure, Youth characters) reveals a consistent pattern — The Simpsons is stably positive while Family Guy starts more negative then converges upward over time, suggesting narrative maturation.

IMDb ratings: no meaningful correlation with per-episode sentiment in either show; ratings show a declining trend with season number, consistent with audience fatigue.

Key Findings

- **Language divergence is empirically confirmed:** Family Guy is significantly more provocative; The Simpsons more accessible and stable.
- **Character centrality shifts over time:** Bart → Homer in The Simpsons; Stewie gains prominence in Family Guy's later seasons.
- **Shared thematic universe, different register:** both shows cover similar macro-topics but treat them with markedly different tone.
- **STM detects nonlinear temporal effects** not captured by simple year regression; production era matters more than political administration.
- **Sliding-window co-occurrences** reveal a cross-show episode: Family Guy's "beer" associations include Simpsons characters (S13E01 crossover).

Technologies

Language: R — `tidytext`, `topicmodels`, `stm`, `quanteda`, `textstem`, `tm`, `ggplot2`, `ggwordcloud`, `patchwork`, `LDAvis`.
Data: Springfield!Springfield! (scripts), IMDb (ratings & synopses).

Future work: intra-episode segmentation for finer-grained analysis; word embeddings for semantic similarity; extension to other long-running shows.