Comparing Probabilistic and Embedding-Based Topic Models for Sentiment-Rich Text: Evidence from IMDB Movie Reviews

Sattiki Ganguly

December 29, 2025

Abstract

Topic modeling is a common approach for uncovering latent thematic structure in large text corpora. Traditional probabilistic models such as Latent Dirichlet Allocation (LDA) rely on bag-of-words representations and word co-occurrence patterns, while modern approaches leverage contextual embeddings derived from large language models such as BERT. This paper compares LDA and BERT-based topic modeling (BERTopic) on a sentiment-rich corpus of movie reviews. Using the IMDB dataset, we evaluate topic coherence and sentiment alignment by examining how topics distribute across positive and negative reviews. Our results show that while LDA captures broad thematic structure, BERT-based topics exhibit stronger semantic coherence and clearer alignment with sentiment polarity. The implementation was performed using Python 3.9. The complete source code, including the preprocessing pipeline, model hyperparameters, and the LaTeX generation scripts, is publicly archived at `https://github.com/Sattiki/nlp-final-research-note`

1 Introduction

User-generated text data plays a central role in understanding public opinion, preferences, and evaluative judgments. Movie reviews, in particular, provide rich subjective narratives that combine emotional expression with descriptive commentary. The IMDB reviews dataset provides an extensive user generated content that has been sufficiently pre-processed to understand the binary sentiment. Reviews have been classified in the binary of either **positive** or **negative** reviews.

This polarity makes it a unique challenge to identify the underlying sentiment of the review. Traditionally Latent Dirichlet Allocation has been used as a bayesian approach to topic modelling. LDA is an unsupervised learning technique that helps uncover latent themes from documents. The goal of LDA has typically been to uncover latent themes and their associated word distributions.

The challenge of LDA is that the model must not only identify the topic (the what) but also the sentiment (the how). Procedurally, LDA ignores order of the word and context much like other topic modelling approaches such as LSA.

LDA operates on a generative probabilistic framework, assuming that documents are mixtures of topics and topics are mixtures of words.

However, LDA is a "Bag-of-Words" (BoW) model, meaning it relies on simplifying assumptions such as word exchangeability. It views text as a statistical distribution of tokens, which limits its ability to capture:

- Semantic Nuance: It cannot distinguish between the different meanings of the same word (polysemy).

- Negation: It often misses how a single word like "not" can flip the entire meaning of a sentence.
- Contextual Meaning: It ignores the order of words, which is where much of human emotion is encoded.

Additionally, LDA lacks from the derivation of the "ground truth". Ground truth acts as a benchmark. Ground truth provides accurately labelled verified information needed to train supervised ML models, validate their performance and test their ability to generalize.
To address these limitations, the field has shifted toward representation learning. Bidirectional Encoder Representations from Transformers (BERT) represents a paradigm shift. Unlike the sparse, frequency-based matrices of LDA, BERT generates dense, high-dimensional contextual embeddings. By utilizing a "Self-Attention" mechanism, BERT encodes the relationship between every word in a sentence simultaneously, capturing the subtle syntactic and semantic structures that signify human sentiment.
Research Objective: This paper investigates the performance gap between classical probabilistic topic modeling and modern transformer-based architectures when applied to sentiment-rich text. Using the IMDb Reviews Dataset as a benchmark, we conduct a comparative analysis to determine if topics extracted via BERT-based semantic clustering align more coherently with sentiment polarity than the statistical clusters produced by LDA. We specifically evaluate whether the "semantic depth" of BERT allows for a more granular understanding of why certain themes (e.g., "Screenwriting" or "Cinematography") drive positive or negative evaluations compared to the "keyword-heavy" approach of LDA.
Research Question: How do classical probabilistic topic models (LDA) compare with modern transformer-based architectures (BERTopic) in their ability to extract semantically coherent and sentiment-aligned themes from subjective datasets like IMDb movie reviews?

2 Dataset and Pre-processing

Sampling - Stratification and Feature Engineering

- The analysis uses the IMDB Dataset of 50,000 movie reviews, evenly split between positive and negative sentiment labels. To reduce computational complexity while maintaining balance, a stratified subsample of 25,000 reviews was constructed, ensuring equal representation of both sentiment classes.
- The dataset was further divided into training and test sets using stratified sampling. All exploratory analysis and topic modeling were conducted on the training set to avoid information leakage.
- Minimal preprocessing was applied. Reviews were converted to strings, and English stop words were removed during vectorization. No aggressive stemming or lemmatization was performed, particularly for BERT-based models, where contextual embeddings already encode semantic similarity.

3 Exploratory Data Analysis

Summary Statistics

Character Level Distribution Initial exploratory data analysis revealed that the mean character length at 1305.914300 was significantly larger than the median character length of 969. Most of the characters lie on the 75th percentile. From this we understand that most reviews are approx less than 1000 characters but there are significant amount of reviews that are long enough that there is a "long tail" of extremely long reviews that are pulling the average upward. This shows a skewness towards the right in the distribution.
Token Level Distribution
In the token level distribution, we use the lambda split to perform a quick, on-the-fly "Token Count" (word count) for each review. This is helpful since LDA only values words as it is a bag of words model.On the other hand the BERT transformer has a hard limit of 512 tokens. A word-based count is the closest approximation to BERT's internal tokenizer. With this we can see yet again the mean of 230.50 is higher than the median 173 and most of the tokens lie on the 75th percentile yet again of 281. This again shows a right skewness in the distribution.
Exploratory analysis reveals substantial variation in the length of the review, with both character-level and token-level distributions exhibiting strong right-hand skewness. A small number of reviews contain several thousand tokens, while most reviews are considerably shorter. Importantly, review length distributions are similar across positive and negative sentiments suggesting that sentiment polarity cannot be inferred from structural features alone.
The dataset is well balanced by to eliminate concerns about class imbalances.

4 Latent Dirichlet Allocation

4.1 Model Specification

Latent Dirichlet Allocation was implemented via CountVectorizer. The count vectorizer turns the text into a matrix of token counts called Bag-of-words, extremely common words that appear more than 95% of the time were removed, extremely rare words that appear just less than 10% were also removed including common english stop words.
With hyperparameter tuning, we tell the model to find 20 distinct themes/topics across the dataset. We initialize the model with batch learning for a maximum iteration of 10 times and then we develop a document-topic matrix.
The document topic matrix is a probabilistic representation of the corpus. While the raw text is high-dimensional and messy, this matrix reduces each review into a concise 20-dimensional vector.Each row represents a unique review and each column a theme.
The topic dataframe is then constructed that uses the document topic matrix. Here, we can preview the distribution of probabilities across the 20 dimensional vector.

4.2 Topic Interpretation

After we construct the topic matrix, we create a function to extract the most important words for each topic
Inspection of the most probable words per topic reveals interpretable themes related to acting quality, narrative structure, genre-specific language, and emotional reactions. However, many topics mix evaluative and descriptive language, reflecting LDA's reliance on word co-occurrence rather than semantic meaning.

Table 1: LDA Topic Sentiment Correlation and Polarity Analysis

Topic ID	Thematic Label	Negative Prob.	Positive Prob.	Net Polarity
0	Family & Life Drama	0.0248	0.0550	+0.0302
1	War & Western Series	0.0193	0.0318	+0.0125
2	Casual Viewing/Engagement	0.0154	0.0254	+0.0100
3	Award-Quality Performance	0.0119	0.0326	+0.0207
5	General Criticism (Bad)	0.1419	0.0290	-0.1129
8	Book Adaptations	0.0344	0.0722	+0.0378
14	Action & Character Development	0.0470	0.0935	+0.0465
15	Sci-Fi & Special Effects	0.0280	0.0197	-0.0083
16	Horror & Slasher Plotlines	0.0580	0.0580	0.0000
17	Negative Critique (Acting)	0.1465	0.0267	-0.1198
18	Biographical & Coming-of-Age	0.0414	0.1079	+0.0665
19	General Movie Recommendations	0.124931	0.101440	-0.0234

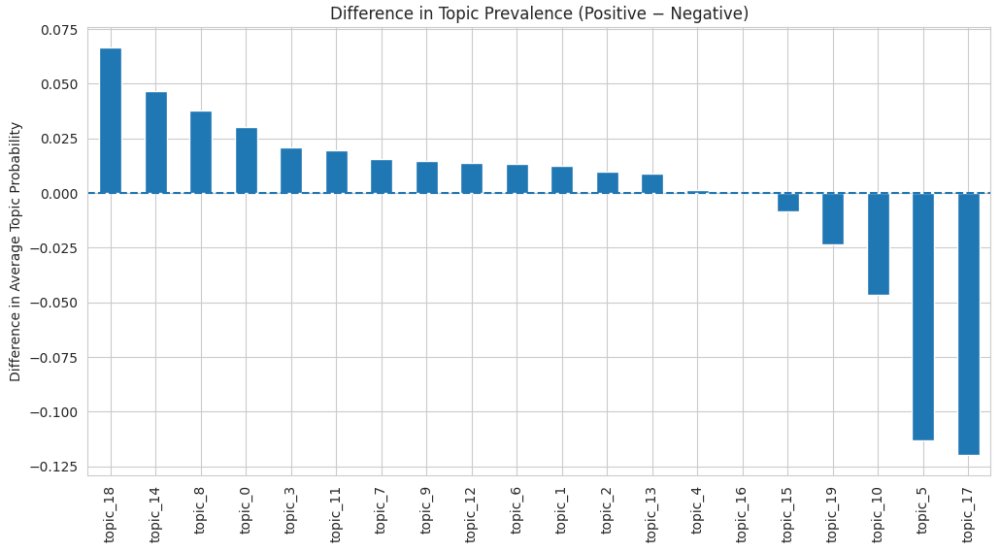

Figure 1: LDA Positive-Negative Prevalence

4.3 Topic–Sentiment Relationship

The keyword extraction from the 20-topic LDA model reveals a clear partitioning of the IMDb corpus into three distinct thematic categories: genre-specific clusters, narrative-driven frameworks, and sentiment-heavy evaluative lexicons. Genre-specific success is most evident in **Topic 15** (Sci-Fi/Special Effects) and **Topic 16** (Horror/Slasher), where technical terms such as *"alien," "monster,"* and *"killer"* appear with high probability. Narrative themes are captured in **Topic 0** and **Topic 13**, which cluster domestic family terms and dance,musical performance vocabulary, respectively. Crucially for this study, **Topic 5** and **Topic 17** emerge as "criticism clusters," where the high frequency of terms like *"bad," "acting,"* and *"don't"* suggests a strong correlation with negative sentiment. However, the consistent presence of the token *"br"* across all topics indicates a common artifact of HTML line-break tags, while the overlap in generic descriptors in **Topics 10, 11,** and **19** illustrates the limitations of a frequency-based model in resolving the semantic boundaries of common evaluative language.

To examine the relationship between topics and sentiment, document–topic distributions were averaged separately for positive and negative reviews. While some topics exhibit higher prevalence in one sentiment class, the magnitude of these differences is generally small. However topic 17 shows a high prevalence of negative reviews which is clear because it deals with negative criticism of acting 1 and 1.
Visualizations of topic polarity, defined as the difference in average topic prevalence between positive and negative reviews, show that most LDA topics cluster near zero. This suggests that LDA captures thematic content but aligns weakly with sentiment polarity.

5 BERT-Based Topic Modeling

5.1 Model Specification

To overcome the limitations of bag-of-words models, we apply BERTopic, an embedding-based topic modeling framework that combines sentence-level BERT embeddings with clustering and class-based TF-IDF representations. We use the all-MiniLM-L6-v2 model which is lightweight but efficient model.
We use the `all-MiniLM-L6-v2` sentence transformer to generate document embeddings. Dimensionality reduction is performed using UMAP, and clustering is carried out with HDBSCAN, allowing the model to identify dense semantic clusters without pre-specifying the number of topics. We extract the labels from the training dataframe to match the topics found and create a a new DataFrame that pairs the Topic ID using the same sentiment classes ie, positive and negative. Furthermore, we remove ouliers. BERTopic uses HDBSCAN for clustering, which identifies 'noise' in the data.
We then group by the 'sentiment' (Positive/Negative) and the 'topic' number assigned by BERT. The resulting table, topic sentiment counts, serves as a real-world test of the model's ability to group meanings. By looking at these numbers, we can distinguish between thematic topics (which appear often and have a balanced mix of likes and dislikes) and evaluative topics (which have skewed numbers and a very strong positive or negative bias).
A high concentration of reviews in just the "positive" or just the "negative" row shows that the BERT model was successful. It proves the model grouped these documents based on the connotative intent (the "feeling" or "vibe") of the reviewer, rather than just the literal denotative subject matter (the basic dictionary definition of the words).

Table 2: Classification of BERT-Generated Topics by Sentiment Distribution

Feature	**Thematic Topics**	**Evaluative Topics**
Description	General subject matter or genre.	Sentiment-driven critiques.
Frequency	Usually high (Commonly discussed).	Variable (Specific to opinion).
Sentiment Bias	Balanced (Mixed Positive/Negative).	Skewed (Polarized Distribution).
Language Type	Denotative (Literal subject).	Connotative (Implied intent/mood).
Examples	Action, Sci-Fi, Setting, Plot.	Masterpiece, Boring, Poor Acting.

5.2 Topic Coherence

BERTopic produces fewer but more semantically coherent topics than LDA. While LDA deals with a bag-of-words BERT uses bidirectional processing. Essentially, BERT reads the vector em-

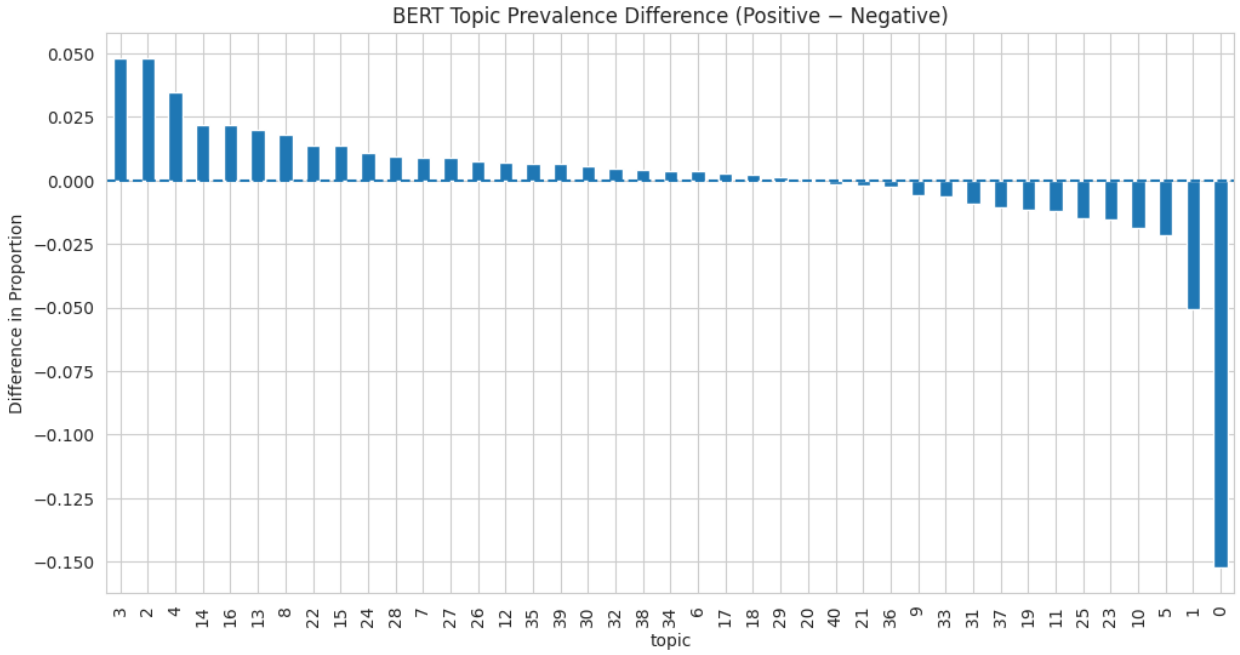

Figure 2: BERTopic Positive-Negative Prevalence

bedding in both directions to understand the "neighborhood" of a word. Due to the self-attention mechanism, the relationship between two words are better attended in BERT.
Top words within topics form meaningful semantic clusters related to emotional engagement, performance quality, narrative strength, and viewer dissatisfaction.
Intertopic distance visualizations reveal well-separated clusters, indicating that embedding-based representations capture high-level semantic distinctions absent from bag-of-words models.

5.3 Topic–Sentiment Relationship

After we set up the dataframe, we see the positive-negative difference in Bert, from the results of the chart we can see the differences in proportion between the proportion of negative to positive. Negative sentiment is more concentrated around topic 0 and positive sentiment is seemingly more well spread out. Topic 0 3 contains more negative feedback as it clusters criticism, hence the sentiment is more negative. Barring topic 0, all other topics distribute sentiment between the two in a similar differential as noted in 2.
Therefore, we see more distribution of polarization under BERTopic rather than LDA.

6 Comparative Analysis

- We observe substantial heterogenity between LDA and BerTopic. It is a matter of label precision and also a result of the self-attention mechanism as discussed earlier present in BERTopic. While LDA relies more on a probabilistic model based on co-occurence. LDA tends to catch broad net and struggles with semantic nuance, negation, and evaluative language
- For instance, when we inspect 1 to see the topic sentiment and thematic labels produced

Table 3: BERTopic Contingency Table: Thematic Labels and Sentiment Proportions

Topic ID	Thematic Label	Negative %	Positive %
0	Negative Evaluative Feedback	84.28%	15.72%
1	TV Shows & Episodic Series	40.16%	59.84%
4	Horror & Slasher Genre	67.07%	32.93%
5	Comedy & Humorous Content	61.59%	38.41%
8	French Language Cinema	55.25%	44.75%
11	Classic Musicals (Kelly/Sinatra)	26.34%	73.66%
12	Western Genre	28.89%	71.11%
13	Vampire & Gothic (Dracula)	66.48%	33.52%
15	Rock Music Documentaries	16.20%	83.80%
19	LGBTQ+ Themes	49.15%	50.85%
20	Zombie & Undead Horror	84.62%	15.38%
23	Christmas & Holiday Specials	59.81%	40.19%
24	Italian Cinema/Neorealism	24.44%	75.56%
25	Shakespeare & Stage Adaptations	35.96%	64.04%
26	Baseball & Sports Drama	32.56%	67.44%
29	Historical Nazi Drama	50.00%	50.00%
30	Dinosaur & Creature Features	73.24%	26.76%
31	Alfred Hitchcock Suspense	35.21%	64.79%
32	Nature Horror (Sharks/Crocs)	81.16%	18.84%
33	Superhero (Batman/Superman)	38.24%	61.76%
35	Kubrick & Intellectual Sci-Fi	36.51%	63.49%

by LDA we find more generic lables such as "Family & Life Drama", "War & Western", "AwardQuality Performances", etc. These labels are helpful but do not provide as much qualititaive insight on the genre of the reviews of the movies. Therefore, it is difficult to truly draw a reasonable insight just by looking at the thematic label created by LDA.

- In contrast, BERTopic leverages transformer based models to produce topics that align more naturally with sentiment. When we inspect the BERTopic Contingency Table 3, we see that there is more granular understanding on the thematic label of the movies. For instance, the labels include for Topic ID 13 "Vampire & Gothic", Topic ID 15 "Rock Music Documentaries", Topic 19 "LGBTQ themes". This shows a far more contexual understanding of the underlying vector embedding compared to the LDA. Qualitiatively, it is far more useful than LDA since we cannot derive much meaningful conclusions from the document-topic matrix compared to the vector-embedding in BERTopic.

- While LDA remains computationally efficient and highly interpretable, BERTopic offers superior semantic coherence and sentiment alignment, making it particularly suitable for subjective, evaluative text. Qualitatively, BERTopic proves far more useful for sentiment auditing. The document-topic matrix in LDA is often "fuzzy," with reviews sharing probabilities across multiple vague themes. However, BERTopic's use of HDBSCAN for clustering allows for the identification of "Dense Sentiment Clusters." For example, the high negative polarity observed in BERT's Topic 32 (Nature Horror) reveals a specific audience disdain for a sub-genre that LDA's broader "Horror" topic completely masked. This suggests that BERTopic does not just find topics; it finds contexts of discussion

7 Limitations

Several limitations should be acknowledged. BERTopic is computationally more expensive and sensitive to hyperparameter choices. Topic models, regardless of approach, do not directly predict sentiment and should not be evaluated as classifiers. Additionally, the analysis is limited to English-language movie reviews and may not generalize to other domains. Furthermore, movie reviews are notoriously known to be sarcastic which is an aspect BERTopic cannot identify well. Furthermore we removed outliers by relegating a large portion of our topics to Topic -1 as outliers in BERTopic. In this process a significant percentage of the user's voice is ignored because it was too unique to form a cluster. BERTopic is also a black box large language model, which doesn't provide much transparency. BERTopic utilizes deep-learning embeddings that operate in high-dimensional vector space. This makes it harder to understand misclassifications where a review might be semantically similar but contextually different. On the other hand, LDA is a much more transparent model as every topic is clearly a distribution over a fixed vocabulary that the researcher can inspect directly via the count vectorizer.

8 Conclusion

This paper demonstrates that while classical probabilistic topic models such as LDA can uncover broad thematic structure, modern embedding-based approaches like BERTopic provide more coherent and sentiment-aligned topics in evaluative text. By comparing topic distributions across sentiment classes, we show that contextual embeddings capture latent semantic structure that aligns closely with human judgments. These findings highlight the value of embedding-based topic modeling for modern text analysis tasks.

9 Appendix

Table 4: Comparative Analysis of LDA and BERTopic Data Pipelines

Pipeline Stage	**LDA (Classical Statistics)**	**BERTopic (Deep Learning)**
Data Input	Bag-of-Words (Sparse Matrix)	Raw Document Strings (Tokens)
Textual Units	Discrete word counts (Tokens)	Contextual Embeddings (Vectors)
Semantic Depth	Shallow (Frequency-based)	Deep (Attention-based mechanism)
Dimensionality	Vocabulary size (V)	384-dimensional vector space
Core Algorithm	Latent Dirichlet Allocation	HDBSCAN Clustering over UMAP
Topic Assignment	Probabilistic (Soft mixture)	Centroid-based (Hard assignment)
Outlier Handling	None (Forced assignment)	Explicit identification (Topic -1)
Key Limitation	Ignores word order and negation	Computationally intensive

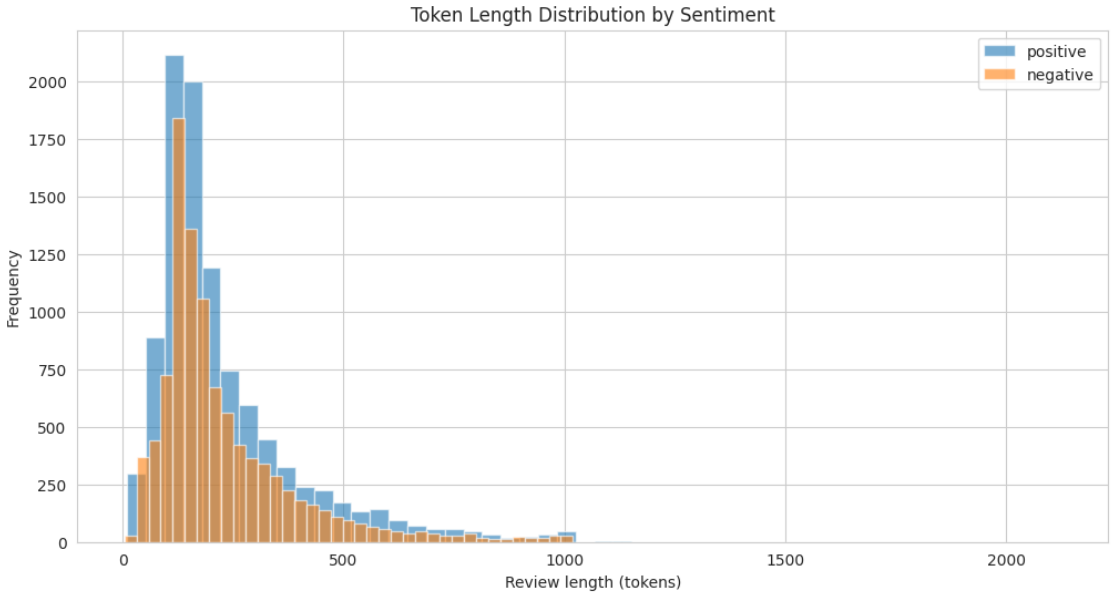

Figure 3: Token Length Distribution by Sentiment (EDA)

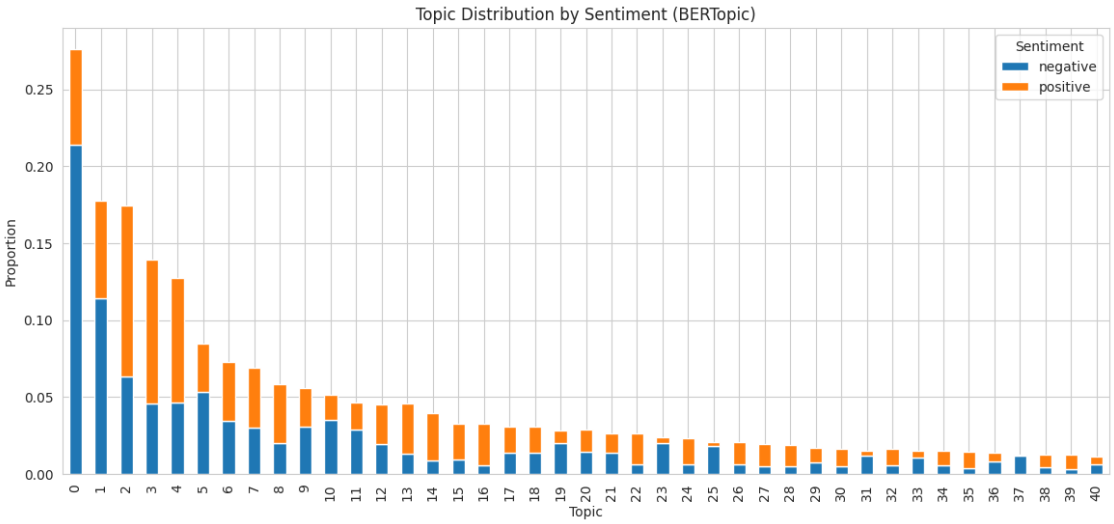

Figure 4: Topic Distribution by Sentiment (BERTopic)

Figure 5: LDA Word Cloud

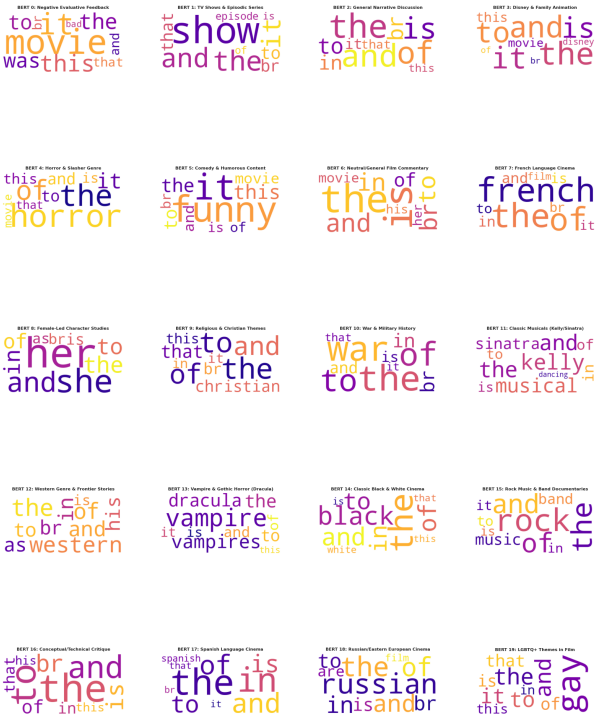

Figure 6: BERTopic Word Cloud

Statement on the Use of AI

Generative AI, namely Google Gemini, was utilized to assist in the pre-processing and visualization tasks of this paper. The AI provided support in debugging Python scripts and providing stylistic guidelines for optimizing LaTeX syntax and table structures. All model interpretations, thematic labeling, and comparative findings were manually verified and finalized by the author to ensure academic integrity and technical accuracy.