Textual Analysis of IMDb Reviews for Zootopia

1. Introduction

With the proliferation of user-generated content in film reviews and social media, understanding the focal points of audience discussions and their modes of emotional expression holds significant implications for film studies, market analysis, and recommendation system development.

Based on IMDb user reviews for Zootopia, this study aims to systematically analyse the content audiences prioritise in their reviews and their evaluative tendencies. Specific research questions include:

(RQ1) Which aspects of the film do audiences primarily discuss in their reviews?

(RQ2) Are reviews concentrated on a few themes or dispersed across multiple aspects?

(RQ3) Are evaluative tendencies in reviews primarily expressed through individual words or phrases?

For RQ1 and RQ2, the LDA topic model was employed to identify latent discussion themes within the comments. Bigram detection was conducted prior to constructing the document-term matrix to capture multi-word phrase information. For RQ3, unigrams, bigrams, and trigrams were extracted and their frequencies counted to analyse whether evaluative expressions were primarily realised through single words or phrases.

2. Data Collection and Cleaning

Data cleaning for RQ1 and RQ2

The overall approach draws upon DataCamp's Topic Modelling tutorial, employing NLTK's English stopword list to filter out low-information functional words, thereby mitigating the impact of high-frequency meaningless terms on LDA topic learning. Secondly, punctuation marks are removed from the text to prevent their misidentification as lexical items, which could otherwise affect bag-of-words models or n-gram statistics. Finally, the WordNetLemmatizer is employed for stemming, unifying different forms of the same lexeme into their base form to reduce lexical sparsity.

Moreover, I implements tailored optimisations based on the specific characteristics of IMDb review data. For instance: merging review titles and body text into a single textual field enables the topic model to leverage both the summarised information from titles and the detailed expressions in the body, thereby more comprehensively reflecting the focal points of audience discussions. Simultaneously, all text undergoes lowercase conversion to ensure identical vocabulary is not misclassified as distinct terms due to differing capitalisation. Moreover, beyond the default stopword list, a customised expanded stopword list was added. This further excluded frequently occurring yet low-distinctiveness words in reviews—such as like, one, really, movie, film, and Zootopia—to enhance the signal-to-noise ratio. Consequently, the LDA topic results became clearer and more comprehensible.

Data cleaning for RQ3

Referencing Analytics Vidhya's text data preprocessing methodology, this study employs regular expressions to remove punctuation, thereby preventing interference with word and phrase statistics; NLTK's English stopword list was employed to remove stopwords, filtering out high-frequency terms with weak semantic information; high-frequency and low-frequency

words were eliminated, removing high-frequency terms with weak evaluative significance and single-occurrence words to reduce noise impact.

Building upon this foundation, this study implemented customised processing according to the analytical requirements of RQ3: unscored comments were removed to ensure each text corresponds to a clear evaluation outcome; Merge titles and body text to extend sentiment analysis coverage to evaluative expressions within comment headings; Unify text to lowercase to prevent identical words being treated as distinct terms due to capitalisation differences; Retain negative words (e.g., `not`, `no`, `nor`) to ensure negation structures remain intact; Preserve evaluative vocabulary (e.g., `good`, `bad`, `love`) to prevent critical sentiment information from being erroneously discarded. Through these treatments, the text data reduces noise while preserving evaluative words and phrases, providing reliable input for subsequent unigram, bigram, and trigram comparative analysis.

This study employs multiple Python libraries for data collection and analysis: Standard libraries (`os`, `time`, `re`, `random`, `hashlib`) for file management, time handling, and text operations; `pandas` for data reading and cleaning; `selenium` for automated web scraping; `nltk` (word segmentation, stopword removal, stemming) and `gensim` (LDA topic modelling and document similarity calculation) for natural language processing, supporting topic analysis and n-gram sentiment orientation research.

Models and Implementation

To address RQ1 and RQ2, this study employs the LDA (Latent Dirichlet Allocation) topic model to analyse review texts for Zootopia on IMDb. The modelling approach draws upon DataCamp's Topic Modelling tutorial and utilises Gensim's LdaModel for implementation. Each review is treated as an independent document, with latent topic structures learned through co-occurrence patterns of terms. During text representation, bigram phrase modelling is introduced. Gensim's Phrases module merges high-frequency co-occurring word pairs into semantically coherent phrases, thereby enhancing the semantic expressiveness and interpretability of the topics. Subsequently, a Bag-of-Words document-term matrix was constructed based on unigrams and bigrams. Low-frequency and high-frequency extreme terms were filtered out as input for the LDA model. The initial number of topics for the LDA model was set to 10, but experiments revealed significant similarity between topics. Consequently, further trials were conducted with 20 and 30 topics. Upon comparison, the results from 20 and 30 topics showed negligible variation. Consequently, 20 topics were selected to ensure thematic distinctiveness while minimising redundancy. The model underwent multiple training iterations utilising automatic hyperparameter learning mechanisms (alpha and eta). It ultimately identified the primary discussion dimensions within audience comments by recognising high-probability keywords associated with each topic, thereby addressing RQ1.

In RQ2, the study examines whether comment discussions centre on a limited number of themes. Upon model training completion, the probability distribution of each film review across all themes is extracted, with the theme exhibiting the highest probability designated as the dominant theme for that review. The number of comments and their respective proportions for each theme are then tallied to characterise the overall topic distribution. Further introducing two centralisation metrics—the Gini coefficient and information entropy—quantitatively assesses the thematic probability distribution: the Gini coefficient measures whether comments cluster around a limited number of topics, while information entropy reflects the dispersion of discussion content. By combining thematic proportion, average probability, and centralisation metrics, we determine whether audience commentary centres on a few core topics or disperses across multiple aspects, thereby providing quantitative evidence for RQ2.

To address RQ3 (which words and fixed expressions audiences commonly employ in reviews to express evaluative tendencies), this study employs n-gram language models to conduct statistical analysis on IMDb review texts. The n-gram model captures local co-occurrence relationships between words by sliding fixed-length windows across text sequences, treating consecutive n words as a single unit. This study constructs unigrams (n=1), bigrams (n=2), and trigrams (n=3) to contrast expression characteristics at word and phrase levels respectively.

At the implementation level, the study utilises the pre-cleaned review text (`clean_text` field). First, NLTK's tokeniser converts each review into a lowercase word sequence. Subsequently, drawing upon Analytics Vidhya's implementation methodology for n-gram principles, a custom `generate_N_grams` function was developed. This function employs list slicing and `zip` operations to generate contiguous word sequences, rather than directly invoking NLTK's built-in `ngrams` interface. This approach more intuitively demonstrates the n-gram construction mechanism—namely, shifting and combining word sequences to form phrases of length n.

The generated unigrams, bigrams, and trigrams were aggregated into a global list, with their frequencies across all comments counted using `Counter`. Ultimately, the top 100 most frequent n-grams were exported as separate CSV files. This enabled subsequent comparative analysis to determine whether audiences favoured single evaluative words or fixed multi-word phrases when expressing opinions in comments. This methodology effectively reveals high-frequency linguistic patterns within comment texts, providing quantitative evidence for understanding audience focal points and evaluative expression styles.

Results

RQ1:

```
Topic_ID,Keywords,Num_of_Reviews,Percentage,Avg_Probability,Gini,Entropy,Concentration_Level
0,"way, character, message, much, animal, judy_hopps",179,18.82,15.36,0.807,92.481,Very High
19,"character, lot, animal, watch, kid, message",92,9.67,10.3,0.865,73.737,Very High
15,"kid, movie, many, character, judy, scene",72,7.57,7.97,0.898,51.665,Very High
6,"character, movie, ever, many, animal, could",56,5.89,6.25,0.92,38.795,Very High
1,"character, zootropolis, kid, way, animated, city",56,5.89,6.06,0.915,46.831,Very High
5,"judy, nick, fox, animal, predator, life",52,5.47,5.42,0.912,53.924,Very High
17,"animated, movie, amazing, message, character, plot",47,4.94,5.15,0.935,27.017,Very High
4,"animal, character, watched, watch, real, entertaining",44,4.63,4.49,0.943,19.989,Very High
10,"kid, thing, movie, funny, year, watch",41,4.31,4.54,0.939,27.946,Very High
16,"character, kid, much, way, city, animal",41,4.31,4.33,0.945,19.115,Very High
18,"people, know, thing, character, everything, another",37,3.89,3.99,0.945,23.683,Very High
13,"character, animal, city, judy, stereotype, way",35,3.68,3.96,0.941,30.04,Very High
3,"character, movie, dont, u, message, animal",33,3.47,3.59,0.949,21.004,Very High
9,"humor, watch, u, plot, character, thing",30,3.15,3.1,0.955,13.731,Very High
11,"character, film, animated, kid, 3d, judy_hopps",29,3.05,3.62,0.948,23.069,Very High
8,"lot, city, kid, funny, friend, line",26,2.73,3.1,0.954,17.123,Very High
14,"humor, animal, judy, bunny, officer, city",24,2.52,2.58,0.957,15.275,Very High
7,"judy, dream, animal, people, city, case",21,2.21,2.24,0.96,12.4,Very High
2,"judy, fox, predator, city, message, rabbit",21,2.21,2.35,0.955,19.84,Very High
12,"message, judy_hopps, mean, perfect, smallest_shrew, largest_elephant",15,1.58,1.6,0.961,9.876,Very High
```

Fig.1 RQ1&RQ2 Result

Audience commentary on this animated film exhibits a high degree of concentration around a few core aspects. Although the model identified multiple themes, these themes show significant overlap at the keyword level, reflecting an overall consistency in audience focus rather than a fragmented discussion structure.

The most prominent focal point of discussion revolves around the characters and their portrayal. Across nearly all themes, ‘character’ emerges as a high-frequency keyword, frequently co-occurring with character-related terms such as ‘animal,’ ‘Judy,’ and ‘Nick.’ This

indicates that audience commentary on the animation predominantly centres on character design, personality, and character relationships. For animated films, characters serve as the primary vehicle for audience emotional resonance, making character-related content a natural focal point in reviews.

Beyond characters, audiences also show significant interest in the themes and social commentary conveyed by the animation. ‘Message’ recurs across multiple topics, forming a semantic cluster alongside terms like ‘people,’ ‘stereotype,’ and ‘life.’ This suggests viewers perceive the animation not merely as entertainment but actively interpret its underlying values and societal implications, indicating widespread discussion of the film's implicit messages and reflections on reality within the reviews.

Furthermore, keywords such as ‘kid,’ ‘funny,’ and ‘watch’ frequently appear in reviews, revealing audience attention to the films' child-oriented attributes and family viewing experience. When evaluating animated films, audiences often consider their suitability for children, their level of entertainment value, and their appropriateness for family viewing. This reflects the close association animated films hold in viewers' minds with family entertainment and children's education.

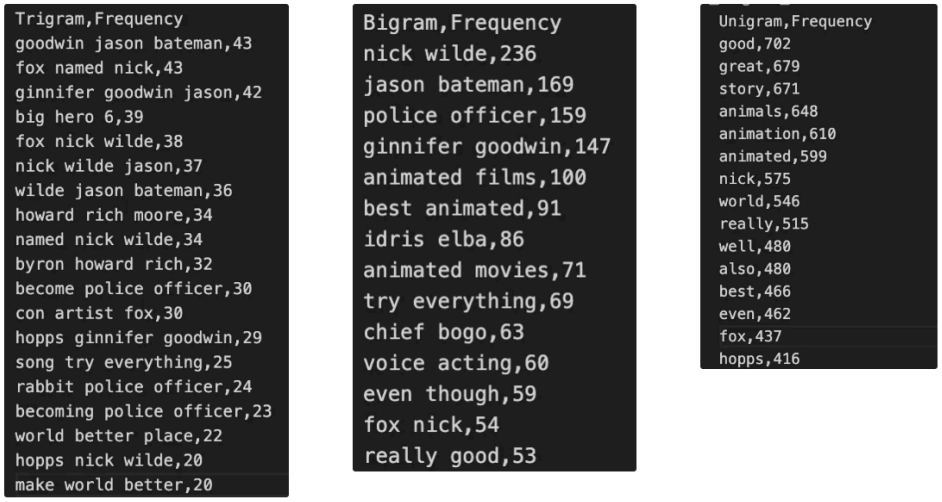
Finally, certain themes address plot structure, comedic impact, and overall viewing experience, as evidenced by keywords like ‘plot,’ ‘humour,’ ‘entertaining,’ and ‘amazing.’ This indicates audiences also evaluate films from perspectives of entertainment value and cinematic immersion. Meanwhile, a minority of reviews mention animation techniques and production features (e.g., ‘animated,’ ‘3D’), though this aspect constitutes a relatively minor proportion of the overall discourse.

Overall, findings for RQ1 indicate that audience commentary on animated films primarily centres on character development, thematic resonance, child-friendliness, and the overall viewing experience. The high similarity between these themes further demonstrates the concentrated and consistent focus of audience attention. This discovery provides crucial contextual grounding for subsequent research into audience emotional expression and evaluative tendencies.

RQ2:

The LDA results reveal substantial overlap in keywords across different themes, with all themes exhibiting extremely high concentration (Gini coefficients exceeding 0.8). This indicates that the model primarily performs granular segmentation within the same semantic space rather than identifying thematically distinct categories. Consequently, in this context, further analysis of whether comments focus on a single theme or multiple themes is unlikely to yield additional explanatory value.

RQ3:



Trigram	Frequency
goodwin jason bateman	43
fox named nick	43
ginnifer goodwin jason	42
big hero 6	39
fox nick wilde	38
nick wilde jason	37
wilde jason bateman	36
howard rich moore	34
named nick wilde	34
byron howard rich	32
become police officer	30
con artist fox	30
hopps ginnifer goodwin	29
song try everything	25
rabbit police officer	24
becoming police officer	23
world better place	22
hopps nick wilde	20
make world better	20

Bigram	Frequency
nick wilde	236
jason bateman	169
police officer	159
ginnifer goodwin	147
animated films	100
best animated	91
idris elba	86
animated movies	71
try everything	69
chief bogo	63
voice acting	60
even though	59
fox nick	54
really good	53

Unigram	Frequency
good	702
great	679
story	671
animals	648
animation	610
animated	599
nick	575
world	546
really	515
well	480
also	480
best	466
even	462
fox	437
hopps	416

Fig. 2. Topic Modeling Results for Unigram, Bigram, and Trigram Features

Through frequency analysis of unigrams, bigrams, and trigrams, the primary methods by which audiences express evaluative tendencies become clearly discernible.

Regarding unigrams, high-frequency words predominantly centre on emotive or evaluative vocabulary, such as good, great, best, fun, amazing, love, funny, well, and really. These words appear with exceptional frequency throughout the review corpus, covering the majority of review texts. This indicates that when expressing attitudes towards the film, viewing experiences, or emotional inclinations, audiences tend to employ direct vocabulary—phrases like ‘good story,’ ‘funny scene,’ or ‘amazing animation’ rely on evaluative words to convey positive sentiment or endorsement directly, demonstrating the efficiency and dominance of the lexical level in conveying emotional information.

At the bigram level, high-frequency phrases include nick wilde, jason bateman, police officer, animated films, best animated, voice acting, and really good. This data reveals bigrams' dual function: firstly, certain phrases contain character names or film-specific terms like ‘nick wilde,’ ‘jason bateman,’ and ‘ginnifer goodwin,’ primarily describing plot content or character interactions while conveying narrative context; On the other hand, certain bigrams (such as ‘really good’, ‘good story’, ‘well done’) explicitly convey evaluative information, combining emotional words with modifiers to form more specific assessments. This indicates bigrams serve both emotional expression and semantic contextualisation within reviews. However, their overall frequency remains lower than high-frequency single words, suggesting a supplementary role in evaluative expression.

Regarding trigrams, high-frequency phrases like ‘goodwin jason bateman’, ‘fox named nick’, ‘big hero 6’, and ‘fox nick wilde’ are almost exclusively character combinations, role names, or film-specific content. Evaluative vocabulary is scarce, and their occurrence is markedly lower than unigrams and bigrams. This indicates trigrams primarily convey plot details, character relationships, or specific events within reviews, rather than directly conveying emotion or evaluative stance. Even where evaluative trigrams exist, their frequency is insufficient to dominate overall sentiment expression.

In summary, audiences primarily rely on single words (unigrams) – particularly high-frequency evaluative vocabulary – to express evaluative stance in film reviews, enabling direct communication of emotion and attitude. Bigrams serve as supplementary tools for capturing evaluative phrases with modifiers (e.g., ‘really good’) or providing contextual details about

characters and plot. Trigrams, conversely, are predominantly used to describe characters, plot developments, or specific events, carrying almost no primary emotional information. Consequently, when conducting sentiment analysis or evaluative tendency research, the most rational strategy is to prioritise unigrams, utilise bigrams as a secondary tool, and reserve trigrams for plot-related semantic analysis.

Evaluation

The application of the LDA topic model assumes that documents are generated from a probability distribution across multiple topics, with each word appearing independently of context ('bag-of-words assumption'). However, for film review texts, words often exhibit contextual dependencies and long-range semantic relationships, such as irony, double negation, or complex emotional expressions. This implies that LDA may fail to capture such nuanced semantics during topic identification, potentially resulting in insufficiently refined topic clustering or the omission of critical information. Furthermore, this study analyses reviews for a single film only, with a sample size of approximately 900 entries, potentially introducing selection bias. When high-frequency words or popular topics dominate, the topic distribution may be skewed by a minority of reviews, thereby compromising clustering efficacy and the accuracy of keyword interpretations. Note that IMDb reviews primarily reflect the views of active users and may not fully represent the broader audience demographic in reality; consequently, the analysis may have certain limitations.

Compared to existing research, this study employs an affective hourglass model and affective lexicon to capture multidimensional affective information, revealing subtle emotional distinctions between films. In contrast, 'Film Review Analysis: Sentiment Analysis of IMDb Movie Reviews' primarily relies on word frequency and phrase co-occurrence analysis. Consequently, its capacity to capture affective dimensions is limited, rendering it incapable of precisely quantifying affective characteristics such as joyfulness or sensitivity. Moreover, despite employing preprocessing and stopword filtering strategies, these methods remain constrained in capturing emotional dimensions. In contrast, this study primarily relies on word frequency and phrase co-occurrence analysis, which has limited capability in capturing emotional dimensions and cannot precisely quantify emotional characteristics such as joy or sensitivity. Moreover, while preprocessing and stopword strategies reduce noise, they may result in the loss of certain emotional information (e.g., insufficient retention of negative vocabulary), thereby compromising the integrity of sentiment analysis. Combining multiple models for comparison—such as sentiment lexicon matching, BERT semantic embeddings, or LDA variants (e.g., Top2Vec, BERTopic)—can enhance thematic interpretability and sentiment analysis accuracy, yielding more robust conclusions. Overall, this approach is suitable for macro-level thematic exploration of review texts, though it exhibits limitations in refined sentiment analysis.

For review analysis, the LDA topic model combined with bigram analysis effectively identified primary audience discussion topics, including character development, plot progression, animation production quality, and social commentary, revealing that reviews concentrate on a few core themes. Concurrently, n-gram analysis indicates that sentiment orientation is primarily expressed through co-occurring words and phrases, with phrases playing a crucial role in conveying complex emotions or evaluations.

These findings suggest that LDA-based thematic analysis assists researchers in summarising discussion focal points across extensive texts, while n-gram analysis complements understanding of sentiment expression, thereby supporting systematic summarisation of audience preferences and discussion content. However, as the model relies on word frequency

and co-occurrence information, subtle sentiments or ironic expressions may be overlooked; thus caution is required when interpreting conclusions regarding the model's limitations. Overall, this methodology successfully addresses the research questions, revealing both the primary concerns expressed by audiences in their reviews and the manner in which these sentiments are articulated. It provides valuable insights for understanding film audience behaviour and designing personalised recommendation strategies.

Conclusions

LDA thematic analysis provides researchers with an effective tool for summarising the focal points of extensive textual discussions, while n-gram analysis aids in deepening our understanding of sentiment expression mechanisms. This collectively supports systematic analysis of audience preferences and discussion content. However, as the model relies on word frequency and co-occurrence information, it exhibits limitations in processing nuanced sentiments or ironic expressions. Consequently, these constraints must be carefully considered when interpreting research conclusions.

Collectively, this research methodology effectively addresses the study's questions, revealing key audience concerns within reviews and their emotional expression patterns. The findings hold multifaceted practical significance and application value:

Film Production Guide: Research indicates that audiences prioritise character development, plot progression and social commentary, offering creative direction for production teams. Producers may refine character design, enhance narrative elements (such as emotional conflict and suspense), and employ nuanced techniques to convey social commentary, thereby resonating with viewers.

Film Marketing Implications: The study's thematic focus provides data-driven support for marketing strategies. Marketing teams can develop targeted promotional plans centred on audience priorities—character appeal, animation quality, and social themes—highlighting these elements in trailers, posters, and social media. By analysing high-frequency vocabulary and emotional expressions, they can design bespoke interactive content and trending hashtags to enhance dissemination impact and user engagement.

Providing an empirical foundation for personalised recommendation systems: The study identifies thematic dimensions and emotional characteristics through LDA and n-gram analysis, directly applicable to recommendation system feature engineering. Incorporating thematic preferences and emotional expression patterns into recommendation algorithms enhances accuracy and user satisfaction, addressing the cold-start and sparsity limitations inherent in collaborative filtering.