Topic Modeling Analysis of Book Reviews: Insights into Themes and Sentiments

Md. Shohanur Rahman Shohan, A.F.M. Rafiul Hassan, and Horish Das Priyo
Department of Computer Science and Engineering
American International University-Bangladesh (AIUB), Dhaka, Bangladesh
{22-46013-1, 22-47048-1, 21-44816-1}@student.aiub.edu

Abstract

This study uses topic modeling of animal ethics and morality book reviews on Goodreads. Through Latent Dirichlet Allocation (LDA) [1], we extract and categorize key themes from user-generated content. We identify 12 main topics: ethical choices, animal rights, pet care and meat eating, among others. These results offer an opportunity to gain insight into the public discourse surrounding animal-related ethical concerns, and suggest situating the findings from topic modeling in a larger conversation about the challenges of analyzing textual data.

Keywords: Topic modeling, LDA, Goodreads reviews, animal ethics, text mining, NLP

1. Introduction

The issue of animal ethics touches on morality, consumption, and rights, making an understanding of public sentiment on the topic important. Book reviews are rich textual data that capture readers' sentiments and reflections. This study utilizes topic modeling based on machine learning [2] to identify the main topics in reviews on Goodreads. Using LDA [1], we find common topics that readers discuss regarding ethical issues with animals.

2. Literature Review

Sentiment analysis and Topic Modeling were used to analyze consumer sentiments and literary topics [3]. It is well established that the book reviews shape public perception and policy debates on ethical issues [4]. LDA circles many uses in natural language processing (NLP) to discover underling concepts in a set of texts [5]. This study extends previous efforts by examining books that eloquently engage with animal ethics and morality.

3. Methodology

3.1 Data Collection

- Web scraping: Goodreads reviews extracted using rvest [6]
- Data format: User-generated reviews in unstructured text form

3.2 Preprocessing

- Text cleaning: Removal of punctuation, numbers, and stopwords
- Conversion to lowercase and stripping of whitespace
- Creation of Document-Term Matrix (DTM) and application of TF-IDF weighting [7]

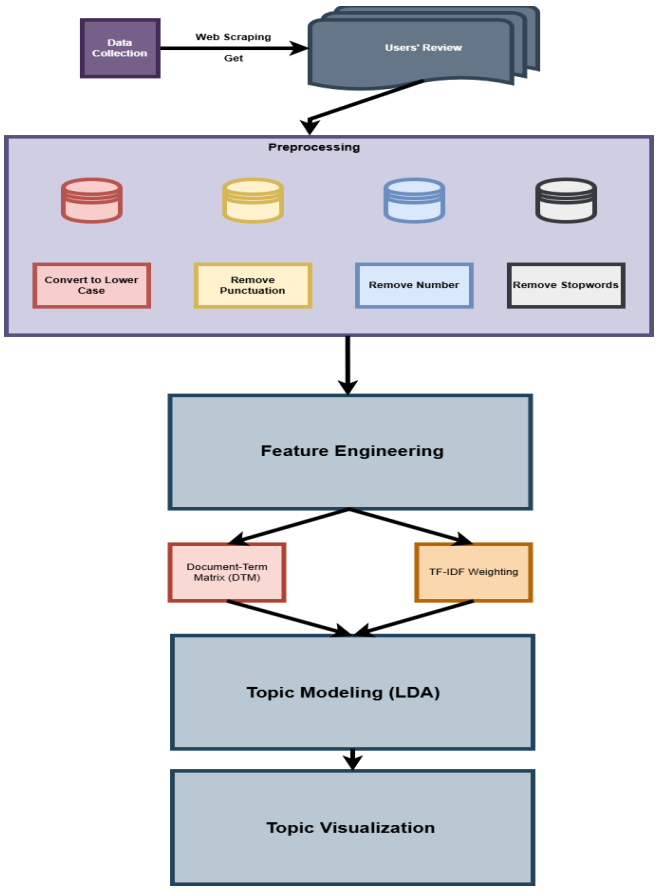
3.3 Topic Modeling

- Implementation of LDA with topicmodels::LDA() [1]
- Number of topics selected: 12 (empirically determined)
- Extraction of top terms for each topic

3.4 Visualization

- Bar plots depicting word probabilities by topic
- Interpretation of thematic clusters

3.5 Workflow Diagram

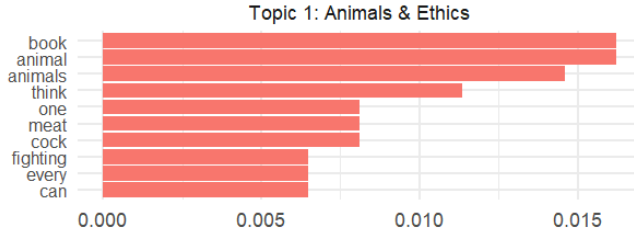


4. Results & Discussion

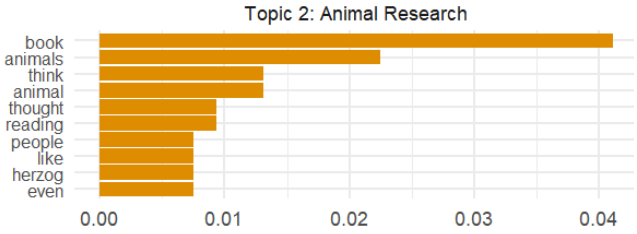
4.1 Identified Topics

The LDA model extracted 12 topics, categorized as follows:

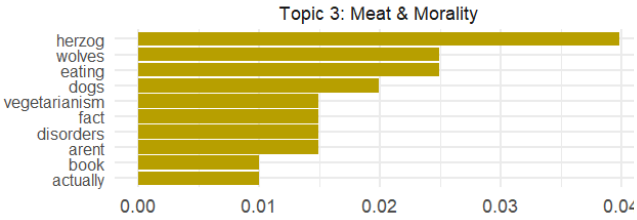
Animals & Ethics: Discussions on the ethical considerations surrounding animal treatment.



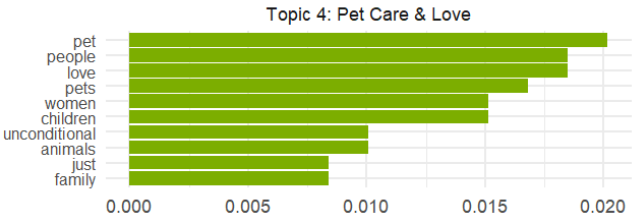
Animal Research: Reviews highlighting studies or opinions on animal-related research.



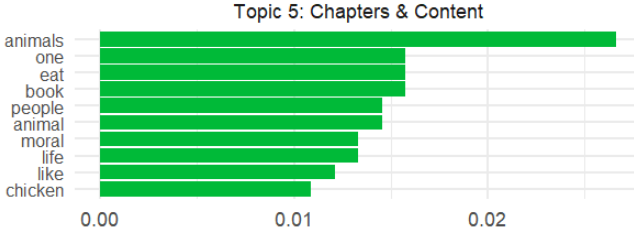
Meat & Morality: Focused on moral dilemmas associated with meat consumption.



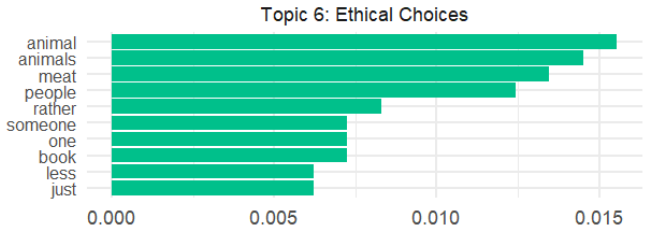
Pet Care & Love: Readers expressing strong emotional bonds with pets and their care.



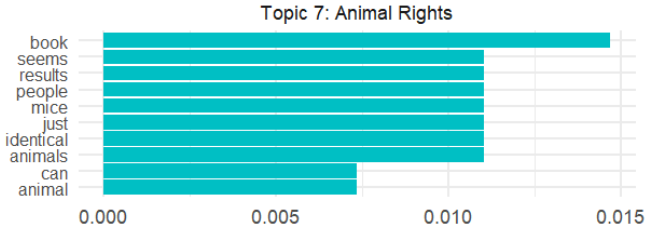
Chapters & Content: Analysis and opinions on specific book chapters and contents.



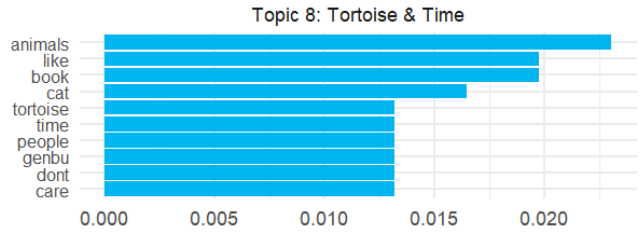
Ethical Choices: Broader discussions on the moral decisions linked to animal rights.



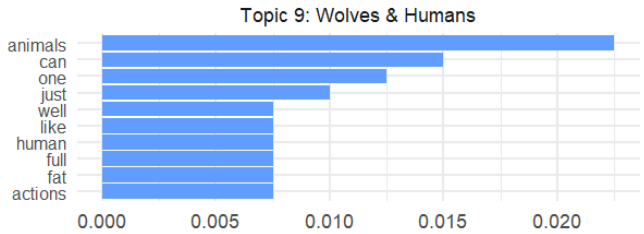
Animal Rights: Advocacy and arguments in favor of protecting animal rights.



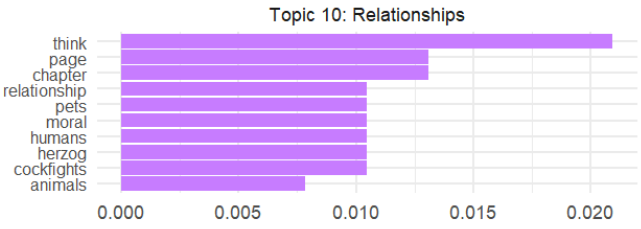
Tortoise & Time: Symbolic or metaphorical references in books involving tortoises and time.



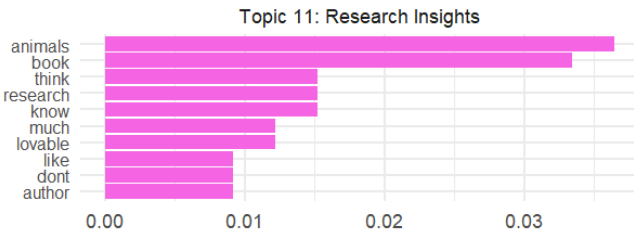
Wolves & Humans: Themes exploring the relationship between wolves and humans.



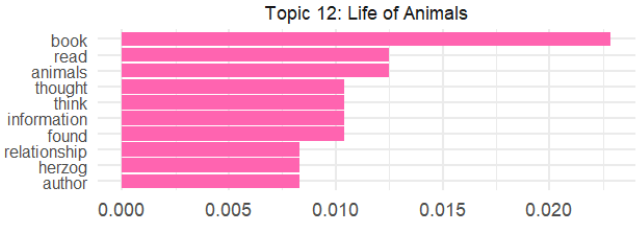
Relationships: Broader discussions on relationships, often tied to animal companionship.



Research Insights: Analytical discussions regarding research methodologies or findings in the books.



Life of Animals: Reflections on the lives and experiences of animals as depicted in the books.



4.2 Interpretation of Findings

- High Probability Terms: Each topic contains key words that define its theme (e.g., "meat," "ethics," "rights").
- Public Discourse on Animal Morality: Reviews often engage in debates about consumption and ethical treatment of animals.
- Reader Engagement: Certain topics, such as "Pet Care & Love," indicate strong emotional connections.

5. CONCLUSION

In this study, we demonstrated the application of LDA for extracting the thematic insights from Goodreads book reviews. The findings offer a glimpse into how readers think about animal ethics, moral dilemmas, and emotional ties with animals. The LDA model uncovered 12 discrete topics related to various dimensions of animal issues, such as pet care, animal rights, and ethical discussions, but it has shortcomings in capturing context-dependent meanings. This illustrates the wealth of diverse themes present in reader-generated content, as well as the larger feature (topic) landscape available for analysis of public discourse surrounding ethical and emotional concepts. Nevertheless, the findings can be further elaborated by conducting advanced techniques such as transformer-based topic model [8] (e.g., BERTopic) and sentiment analysis to grasp readers' sentiments and sentiments. Future research could also compare reader perceptions across different platforms or integrate multimodal data to provide a more holistic view of engagement with literary content.

ACKNOWLEDGMENT

First and foremost, we would like to express our heartfelt gratitude to our supervisor, Tohedul Islam, for his guidance, support, and encouragement throughout the research process. His insight, rigorous critique and persistent inspiration have genuinely enabled the achievement of this study. We are incredibly blessed to have had the opportunity to work under his guidance, we genuinely appreciate all of his efforts to help further both academic growth as well as foster the research embodied in this document.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, no. Suppl 1, pp. 5228-5235, 2004.

[3] K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002. [Online]. Available: http://mallet.cs.umass.edu

[4] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262-272, 2011.

[5] Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[6] H. Wickham, "rvest: Easily harvest (scrape) web pages," Journal of Open Source Software, vol. 1, no. 1, p. 3, 2016.

[7] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning, pp. 133-142, 2003.

[8] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.

APPENDIX

This appendix contains the R code used in the project for web scraping, text preprocessing, topic modeling, and data visualization. The code serves as a reference for reproducing the results discussed in the report and demonstrates the methodologies applied.

A.1. Libraries Used

```
library(rvest)
library(dplyr)
library(tm)
library(topicmodels)
library(ggplot2)
library(tidytext)
```

A.2. Web Scraping Goodreads Book Reviews

```
book_review_url <- "https://www.goodreads.com/book/show/6953508-some-we-love-some-we-hate-some-we-eat"
book_webpage <- read_html(book_review_url)

book_reviews <- book_webpage %>%
  html_elements("#ReviewsSection .Formatted") %>%
  html_text()

book_reviews_df <- data.frame(reviews = book_reviews, stringsAsFactors = FALSE)
book_reviews_df
```

A.3. Text Preprocessing

```
book_corpus <- Corpus(VectorSource(book_reviews_df$reviews))

book_corpus <- tm_map(book_corpus, content_transformer(tolower))
book_corpus <- tm_map(book_corpus, removePunctuation)
book_corpus <- tm_map(book_corpus, removeNumbers)
book_corpus <- tm_map(book_corpus, removeWords, stopwords("en"))
book_corpus <- tm_map(book_corpus, stripWhitespace)

book_reviews_cleaned_df <- data.frame(updated_book_reviews = sapply(book_corpus, as.character), stringsAsFactors = FALSE)
book_reviews_cleaned_df

write.csv(book_reviews_cleaned_df, "D:/Final-Data-Science-Project/Group_10_Final_Term_Project_Dataset.csv", row.names = FALSE)
```

A.4. Feature Engineering with Document-Term Matrix (DTM) and TF-IDF

```
book_dtm <- DocumentTermMatrix(book_corpus)
book_dtm_matrix <- as.matrix(book_dtm)
print(book_dtm_matrix[1:5, 1:5])

book_tfidf <- weightTfIdf(book_dtm)
book_tfidf_matrix <- as.matrix(book_tfidf)
head(book_tfidf_matrix[1:5, 1:5])
```

A.5. Topic Modeling Using Latent Dirichlet Allocation (LDA)

```
num_topics <- 12
book_lda_model <- LDA(book_dtm, k = num_topics, control = list(seed = 1234))
book_term_probs <- tidy(book_lda_model)
book_term_probs


book_top_terms <- book_term_probs %>%
  group_by(topic) %>%
  arrange(desc(beta)) %>%
  slice_head(n = 10) %>%
  ungroup() %>%
  mutate(term = reorder_within(term, beta, topic))
```

A.6. Assigning Topic Labels

```
topic_labels <- c(
  "Topic 1: Animals & Ethics",
  "Topic 2: Animal Research",
  "Topic 3: Meat & Morality",
  "Topic 4: Pet Care & Love",
  "Topic 5: Chapters & Content",
  "Topic 6: Ethical Choices",
  "Topic 7: Animal Rights",
  "Topic 8: Tortoise & Time",
  "Topic 9: Wolves & Humans",
  "Topic 10: Relationships",
  "Topic 11: Research Insights",
  "Topic 12: Life of Animals"
)

book_top_terms <- book_top_terms %>%
  mutate(topic_label = topic_labels[topic])
```

A.7. A.7. Visualization of Topic Modeling Results

```
ggplot(book_top_terms, aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic_label, scales = "free", ncol = 3) +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Word Probabilities by Topic (Book Reviews)",
       x = "Keywords/Terms",
       y = "Probability (Beta)") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8))
```