GoodReads Reviews Summarization

David Nicolay, Kellen Mossner, Matthew Holm

Department of Statistics and Actuarial Science
Stellenbosch University
{26296918, 26024284, 26067404}@sun.ac.za

Abstract

Our analysis uses various natural language processing (NLP) techniques to analyze book reviews scraped from GoodReads. We develop a method to automate the extraction of insights, generate summaries, and predict star ratings for book reviews along with positive and negative aspects reported by reviews. We employ techniques such as sentiment analysis, text generation, aspect-based sentiment analysis (ABSA), and summarization, integrating methods from traditional lexicon-based approaches to modern transformer models.

Introduction

The main objectives of our analysis are to extract valuable insights from book reviews and synthesize this information into useful text like summaries, aspect-based review analysis, and star rating predictions. These methods can help automate tasks such as generating content for publishers, understanding reader sentiment, and deriving meaningful insights from a large collection of text data.

We explore several NLP methods learned in class, including:

- **Scraping**: Using R and Python to collect data ethically.
- **Preprocessing**: Cleaning and structuring the dataset.
- **Exploratory Data Analysis (EDA)**: Exploring relationships and distributions of predictors.
- **Sentiment Analysis**: Classifying reviews as positive, neutral, or negative.
- **Summarization**: Generating concise summaries for individual reviews and aggregating the final summary.
- **Aspect-Based Sentiment Analysis (ABSA)**: Analyzing the sentiment of specific book aspects such as plot or character development.

Ideas and Application

Our project aims to automate the analysis and synthesis of book reviews, using the following key techniques:

- **Sentiment Analysis**: Understanding the overall sentiment of reviews.
- **Summarization**: Using transformer models to summarize individual reviews and generate an overall summary.
- **Book Description Generation**: Leveraging text generation models like BART to create engaging descriptions of books.
- **Star Rating Prediction**: Predicting star ratings using a neural network model trained on review text.
- **Aspect-Based Sentiment Analysis (ABSA)**: Focusing on specific aspects of reviews to provide more nuanced insights.

Implementation

We divide our implementation into stages:

Web Scraper

First, we developed a book web scraper to get the links to a large number of books to be used for further scraping. With this data, we scraped 120 reader reviews from each book using Selenium. Along with the review text, we collected the star rating, review likes and review date. Separately, the genres of the books were also scraped.

Sentiment Analysis

We implement lexicon-based sentiment analysis using NRC and Bing lexicons in R, detecting correlations between sentiment scores and star ratings.

Summarization of Reviews

Individual reviews are summarized using the `facebook/bart-large-cnn` model, with additional fine-tuning using the Sentence-BERT encoder-decoder approach. These models effectively condense reviews into concise versions.

Book Description Generation

We use the BART (Lewis et al. 2019) text generation model with prompts to create book descriptions. Prompt engineering was necessary to achieve engaging and objective descriptions.

Star Rating Prediction

We employ a sequential neural network implemented in TensorFlow/Keras, predicting star ratings based on review text.

Aspect-Based Sentiment Analysis (ABSA)

For ABSA, we used the InstructABSA model based on Tk-Instruct and T5, which provides sentiment analysis on specific aspects such as plot, character, and writing style.

BART Architecture

BART is a transformer-based sequence-to-sequence model introduced by Facebook AI, combining the advantages of bidirectional and auto-regressive models. The model uses an encoder-decoder architecture: the encoder processes the entire input (like BERT), and the decoder generates output token-by-token (like GPT).

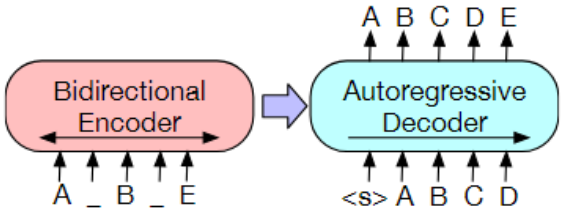


Figure 1: BART Overview

Pretrained on various denoising tasks to learn language patterns, BART was fine-tuned specifically on the CNN/-DailyMail dataset to improve its summarization capabilities. During fine-tuning, the decoder attends to the encoder's output to generate contextually appropriate summaries (Attention Mechanisms). Beam search and length penalty were also used to avoid overly short summaries.

Sentence-BERT

SBERT (Reimers and Gurevych 2019) modifies BERT's architecture in the following ways:

- Employs siamese network structure with twin BERT models sharing identical weights
- Removes BERT's classification head
- Processes sentence pairs in parallel through the twin networks

Technical Summary of InstructABSA

This model introduces positive, negative, and neutral examples to each training sample, and instruction tunes the model Tk-Instruct for ABSA subtasks.

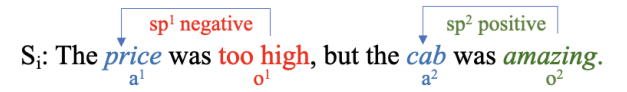
The ABSA subtasks can be represented as follows: Let S_i represent the i^{th} review sentence in the training sample, where $S_i = \{w_i^1, w_i^2, ..., w_i^n\}$ with n as the number of tokens in the sentence.

Each S_i contains a set of aspect terms denoted by $A_i = \{a_i^1, a_i^2, ..., a_i^m\} | m \leq n$, and the corresponding opinion terms and sentiment polarities for each aspect term are denoted by $O_i = \{o_i^1, o_i^2, ..., o_i^m\}$ and $SP_i = \{sp_i^1, sp_i^2, ..., sp_i^m\}$ respectively, where $sp_i^k \in [positive, negative, neutral]$.

The ABSA tasks are described as follows:

ATE: $A_i = LM_{ATE}(S_i)$
$ATSC : sp_i^k = LM_{ATSC}(S_i, a_i^k)$
$ASPE : [A_i, SP_i] = LM_{ASPE}(S_i)$
$AOOE : o_i^k = LM_{AOOE}(S_i, a_i^k)$
$AOPE : [A_i, O_i] = LM_{AOPE}(S_i)$
$AOSTE : [A_i, O_i, SP_i] = LM_{AOSTE}(S_i)$

In these equations, LM represents the language model, and the corresponding inputs and outputs are defined accordingly. InstructABSA instruction tuned $LM_{subtask}$ by prepending task-specific prompts to each input sample to arrive at $LM_{subtask}^{Inst}$



Subtask	Input	Output
Aspect Term Extraction (ATE)	S_i	a^1, a^2
Aspect Term Sentiment Classification (ATSC)	$S_i + a^1$, $S_i + a^2$	sp^1, sp^2
Aspect Sentiment Pair Extraction (ASPE)	S_i	(a^1, sp^1), (a^2, sp^2)
Aspect Oriented Opinion Extraction (AOOE)	$S_i + a^1$, $S_i + a^2$	o^1, o^2
Aspect Opinion Pair Extraction (AOPE)	S_i	(a^1, o^1), (a^2, o^2)
Aspect Opinion Sentiment Triplet Extraction (AOSTE)	S_i	(a^1, o^1, sp^1), (a^2, o^2, sp^2)

Figure 2: ABSA Subtasks Overview (Scaria et al. 2023)

Tk-INSTRUCT

Tk-INSTRUCT is a transformer model trained to follow various in-context instructions. It builds on the T5 text-to-text transformer model using an instruction tuning approach. It converts diverse NLP tasks into a consistent instruction format through:

- **Definition:** Task description
- **Things to avoid:** Common mistakes
- **Positive examples:** Good completions
- **Negative examples:** Poor completions
- **Input**
- **Output**

Understanding the T5 Architecture

T5 (Text-To-Text Transfer Transformer) was proposed by Google in 2020. It was trained on a cleaned common crawl web extracted text corpus. The model uses an unsupervised objective where words are dropped out independently and replaced with sentinel tokens.

Input Representation

1. **Tokenization:** Uses SentencePiece to create a vocabulary of subword units.

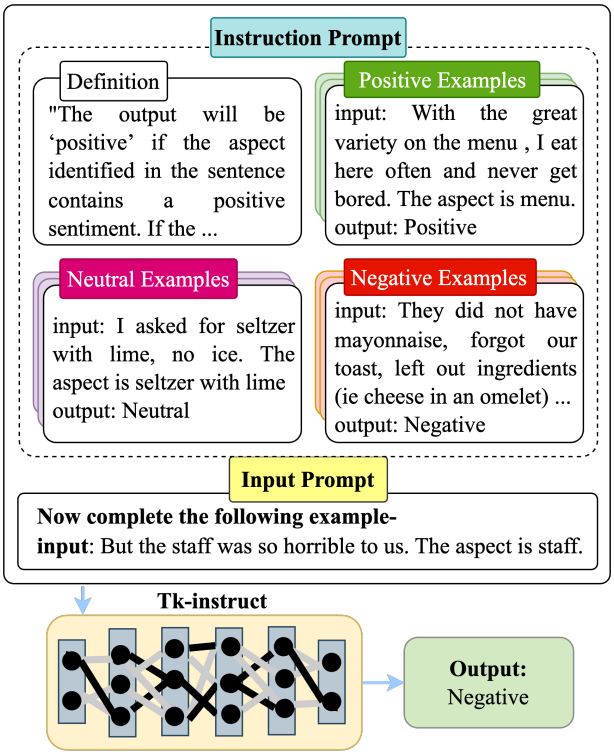


Figure 3: InstructABSA Architecture Overview (Scaria et al. 2023)

2. **Conversion to Token IDs:** Maps tokens to unique integer IDs.
3. **Embedding:** Converts token IDs to dense vector embeddings with positional information.

Encoder The encoder contains multiple layers with:

1. **Self-Attention Mechanism (Vaswani et al. 2023):** Transforms input text into query (Q), key (K) and value (V) vectors:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

2. **Feed-Forward Neural Network:** 2-layered network with ReLU activation
3. **Layer Normalization:** Applied pre-norm without scaling and bias parameters
4. **Residual Connections:** Skip-connections around each sub-layer

Decoder The decoder structure mirrors the encoder with additional:

- Masked self-attention to prevent attending to future positions
- Cross-attention mechanism enabling decoder positions to attend to encoder outputs

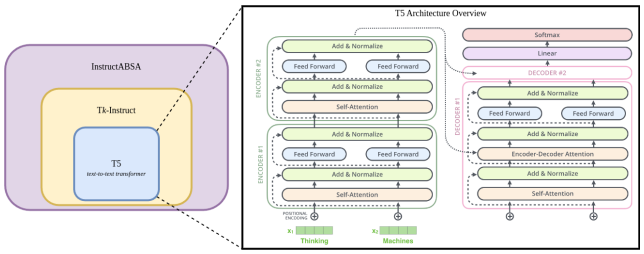


Figure 4: T5 Architecture Overview

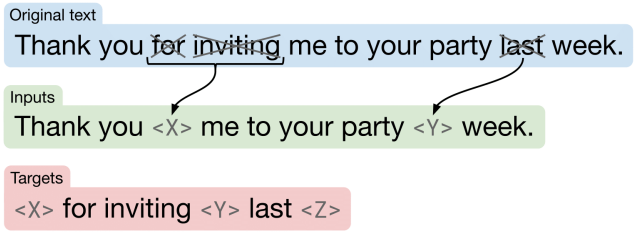


Figure 5: T5 Training Objective (Raffel et al. 2023)

Results and Findings

Sentiment Analysis

Most reviews tended to be positive, reflecting that users typically review books they enjoyed. However, some genres like horror had more polarized sentiments.

Summarization

The summarization models effectively captured the key points of reviews, distilling long reviews into concise summaries that preserved meaning.

Book Descriptions

The generated book descriptions were engaging, though challenges persisted in ensuring the output was consistently written in the third person.

Star Rating Prediction

Our neural network model achieved moderate statistical results, with accuracy improving when shorter reviews were used as input. Through experimentation, we discovered that despite the moderate statistical accuracy, the predictions actually ended up being rather accurate to the human eye. Regularization techniques showed minimal improvement.

Aspect-Based Sentiment Analysis

ABSA provided detailed sentiment analysis for specific aspects of books, allowing us to identify which features of a book (e.g., plot, character development) readers praised or criticized.

Ethical Considerations

- No personal data was scraped.
- Data is used exclusively for demonstrating the NLP and Deep Learning methods we were taught in class.

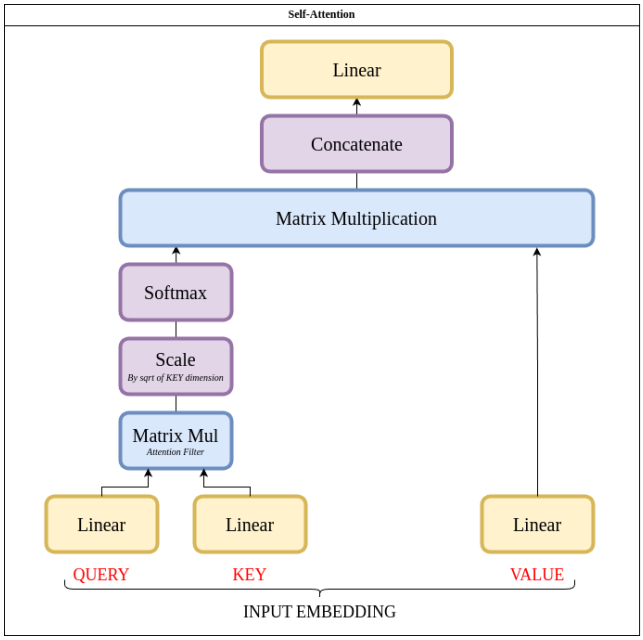


Figure 6: Self-Attention Mechanism

- Our webscrapers are equipped with rate-limiter to ensure an unreasonable rate and amount of traffic is not sent to the GoodReads servers.
- No security measures that were in place to prevent web scrapers from accessing the website were bypassed. We only scraped the data that is publicly available.

Conclusion

Our project demonstrates how NLP techniques can be applied to automate the analysis and synthesis of book reviews. Combining sentiment analysis, ABSA, summarization, and text generation provides valuable insights and aids in content generation, making it useful for publishers, readers, and authors alike.

References

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Scaria, K.; Gupta, H.; Goyal, S.; Sawant, S. A.; Mishra, S.; and Baral, C. 2023. InstructABSA: Instruction Learning for Aspect Based Sentiment Analysis. arXiv:2302.08624.

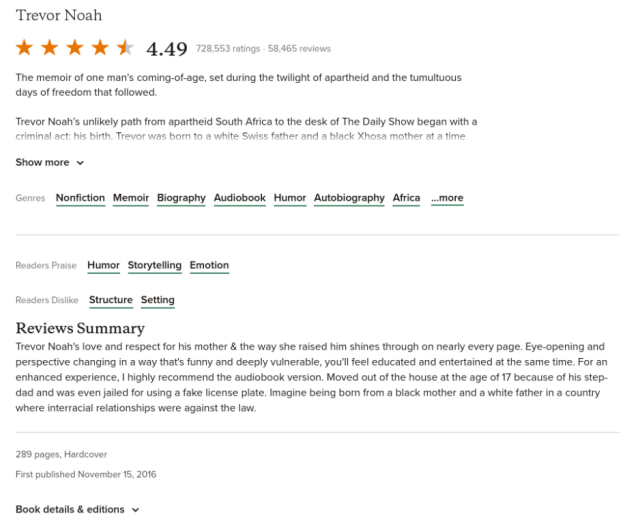


Figure 7: Final product with ABSA and summary sections displayed.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.