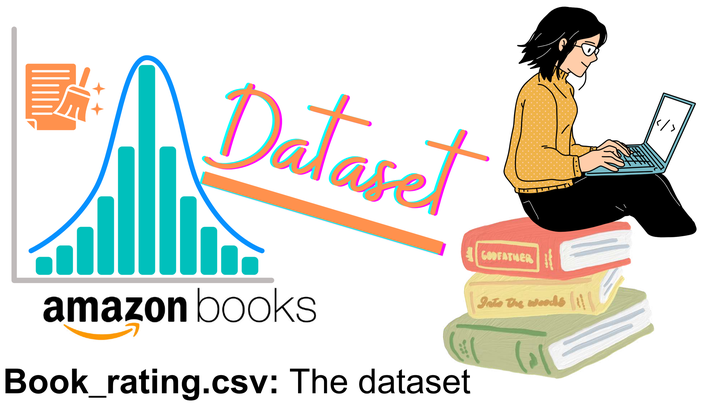
Beyond the Page

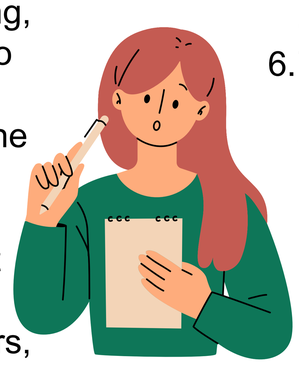
Project Github Link

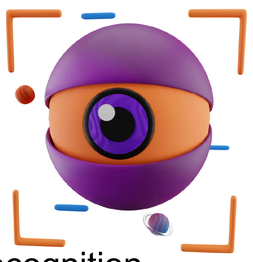
In today's digital age, analyzing readers' sentiments toward books has become crucial for authors, publishers, and the reading community. Online platforms, particularly Amazon, have transformed book sales and amassed a wealth of user reviews. Our project leverages advanced text mining to extract deep insights from these Amazon book reviews, offering an in-depth perspective on the literary landscape.



Book_rating.csv: The dataset consists of 3 million book reviews covering 212.404 unique books. Each entry includes the book's ID, title, price, user ID, user's profile name, review helpfulness rating, review score (ranging from 0 to 5), review time, review summary, and the full text of the review.

Books_data.csv: The dataset contains details of 212.404 books: title, description, authors, cover image URL, preview link, publisher, published date, info link, categories (genres), and average rating count.

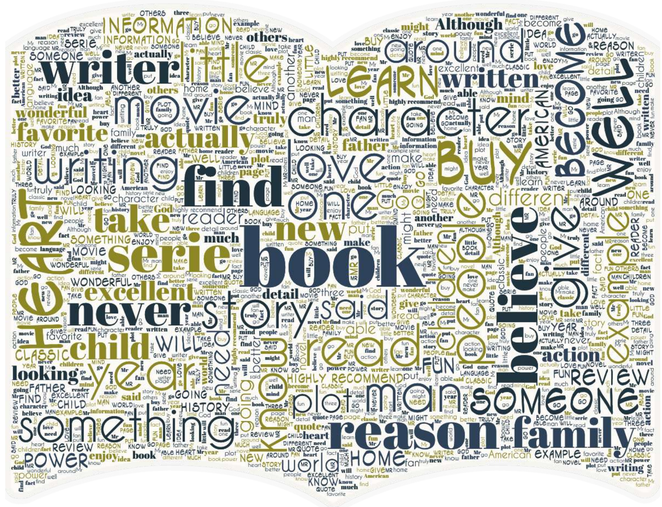




Data Pre-processing Overview

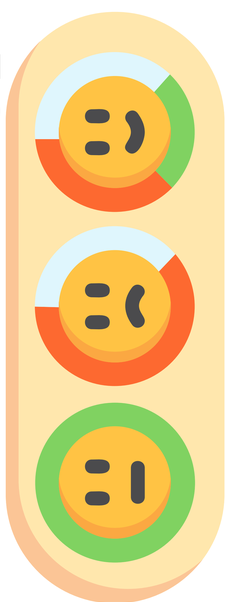
To ensure a solid foundation for our analysis of Amazon book reviews, we streamlined our dataset through the following key steps:

1. Merging: Combined 'Book_rating' and 'Book_data' based on 'Title' to unify relevant data.
2. Column Selection: Extracted essential columns into a new dataframe for focused analysis.
3. Duplicates Removal: Eliminated duplicate entries to maintain data integrity.
4. Missing Values: Removed rows with any missing data to ensure completeness.
5. Data Sampling: Selected a random sample of 20,000 rows, creating a representative dataset data.
6. Text Preprocessing: Used regular expressions for pattern extraction and added a 'word_count' column for 'review/text'.



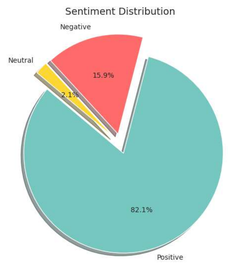
NERC

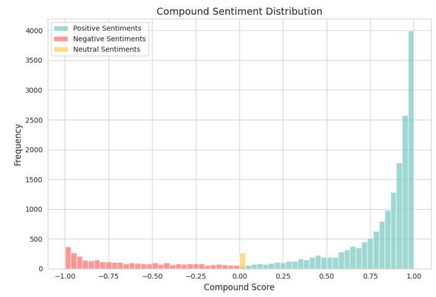
We employed Named Entity Recognition (NER) using spaCy's ‘en_core_web_sm’ model. It reads sentences from a file and assigns BIO (Beginning, Inside, Outside) tags to each token based on the entities recognized by spaCy. The BIO tags are compared with annotated tags from another file using a classification report, where BIO tags are encoded for comparison using `LabelEncoder`. This approach evaluates the performance of NER on the given data, assessing precision, recall, and F1-score for each entity type.

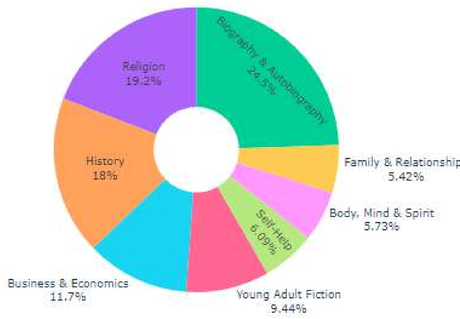


Sentiment Analysis

We applied sentiment analysis to Amazon book reviews using five models: BERT, LSTM, Logistic Regression, Random Forest, and Naive Bayes. Ranked by performance, BERT emerged as the most effective, followed by LSTM, Logistic Regression, Random Forest, and Naive Bayes. This prioritization reflects each model's capability to interpret and classify the sentiments expressed in the reviews, with BERT and LSTM leading due to their advanced contextual understanding and memory of text sequences, respectively.

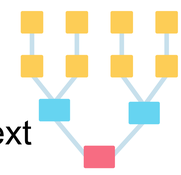






Exploratory Data Analysis

Our EDA focused on dissecting the Amazon Books Reviews dataset to uncover key insights. We analyzed review scores and text content, using statistical summaries and trend analysis to grasp the data's overall landscape. Text analysis revealed common themes and sentiments across reviews, while correlation studies helped identify factors influencing reader engagement. Visual tools like histograms and heatmaps clarified these insights, ensuring our findings were easily accessible. This foundational EDA equipped us with the necessary understanding to delve deeper into NLP applications, setting the stage for targeted analysis and interpretation.



Topic Classification

We conducted topic classification on balanced merged datasets comprising sports tweets, movie reviews, and book reviews (the latter also used in our sentiment analysis). To tackle this diverse dataset, we employed BERT, Multinomial Naive Bayes (MNB), Random Forest, and SVM models. For SVM, Random Forest, and MNB, we integrated TfIDF vectorization and LSA for dimensionality reduction. Performance-wise, BERT stood out as the top performer, demonstrating superior contextual analysis capabilities. It was followed by MNB, Random Forest, and SVM, in that order. This ranking underscores the effectiveness of BERT and MNB in accurately classifying topics from varied text sources, highlighting their robustness in handling complex data.

Division of work:

We collaborated on coding and analysis using Google Colab Jupyter notebooks and prepared our poster with Canva, leveraging both platforms for efficient teamwork and creative design.

- **Adnan**: Code [worked on topic classification and sentiment improvement] & Analysis[topic analysis] & Poster[reviewed the second version and prepared the submission zip file.]
- **Mohammad**: Code [worked on EDA and sentiment analysis] & Analysis [EDA and sentiment analysis] & Poster [created a second and final version of poster, incorporating some text elements from Kiki's version.]
- **Koen**: Code [worked on NERC] & Analysis[NERC analysis] & Poster [reviewed the second version, adding the wordcloud.]
- **Kiki**: Code [worked on NERC] & Analysis[NERC analysis] & Poster [crafted the initial poster design]