Deciphering Sentiments: Machine Learning Insights into Book Reviews

Capstone Project: CIND820

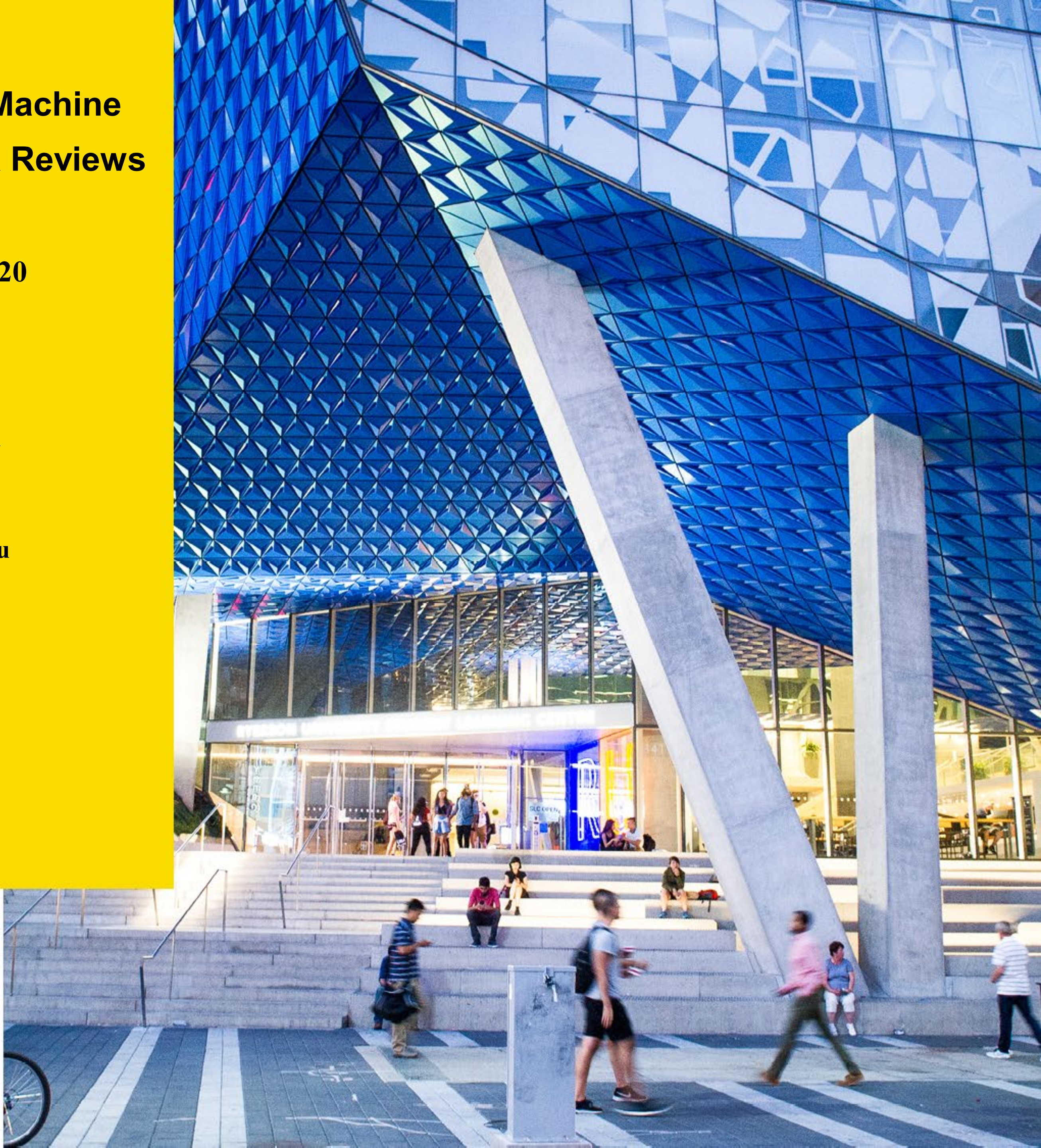
Project by: Tareq HatHat

tareq.hathat@torontomu.ca

Student #501275635

Supervisor: Dr. Ceni Baboglu

Winter-2024

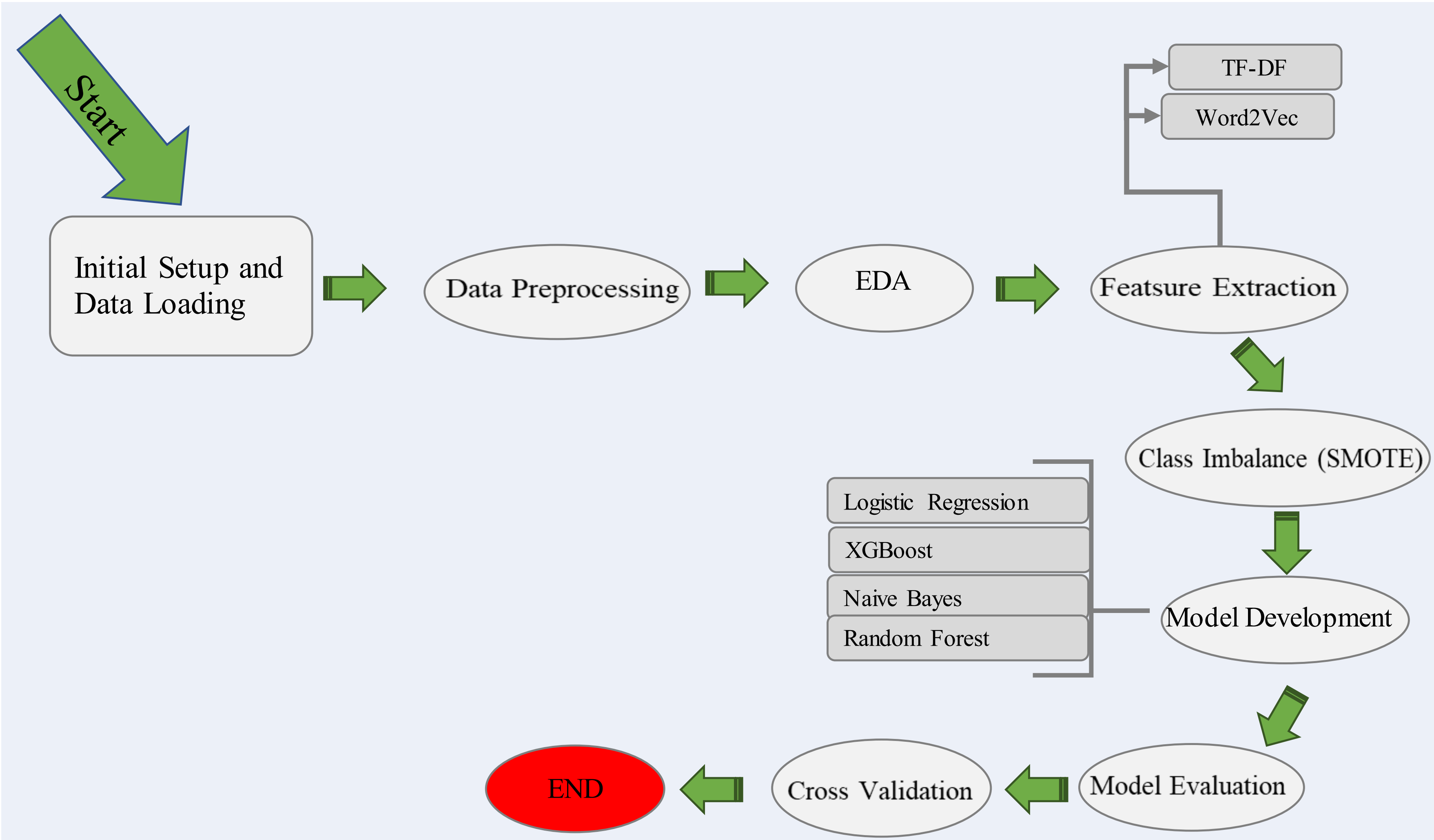


Ryerson University

Project Objectives & Research Questions

- Objective: To understand sentiment distribution within book reviews, focusing on genre/author impact and feature extraction methods' efficacy.
- RQ1: What is the overall sentiment distribution in book reviews?
- RQ2: How do specific genres or authors influence sentiment tendencies?
- RQ3: How effective are TF-IDF and Word2Vec in enhancing sentiment analysis accuracy?

Methodology Overview

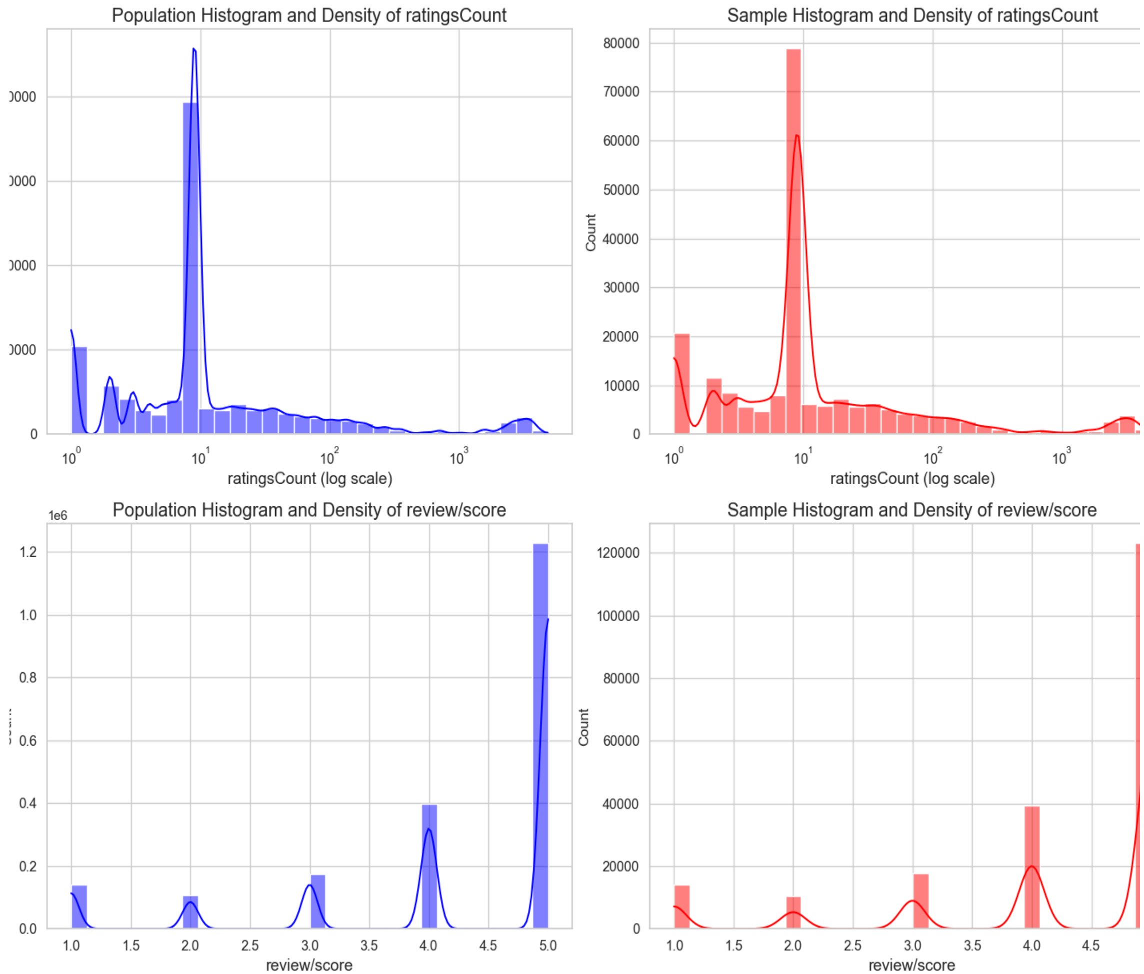


Data Description and Sampling Necessity

- **Dataset Overview**: Utilized two CSV files, each approximately 2.83 GB, encompassing 300,000 rows and 19 combined attributes after merging.
- **Sampling for Analysis**: Due to computational limits, a 10% stratified sample was extracted to maintain sentiment representation.
- **Core Attributes**: The research focused on eight key attributes including Title, Description, Authors, Categories, RatingsCount, Review/Score, Review/Summary, and Review/Text.
- **Statistical Summary**: RatingsCount ranged from 1 to 4,895 with most reviews scoring around 4.22, reflecting positive skewness.
- **Preprocessing Necessity**: Handled missing values, especially in RatingsCount, and removed duplicates to clean data for analysis.

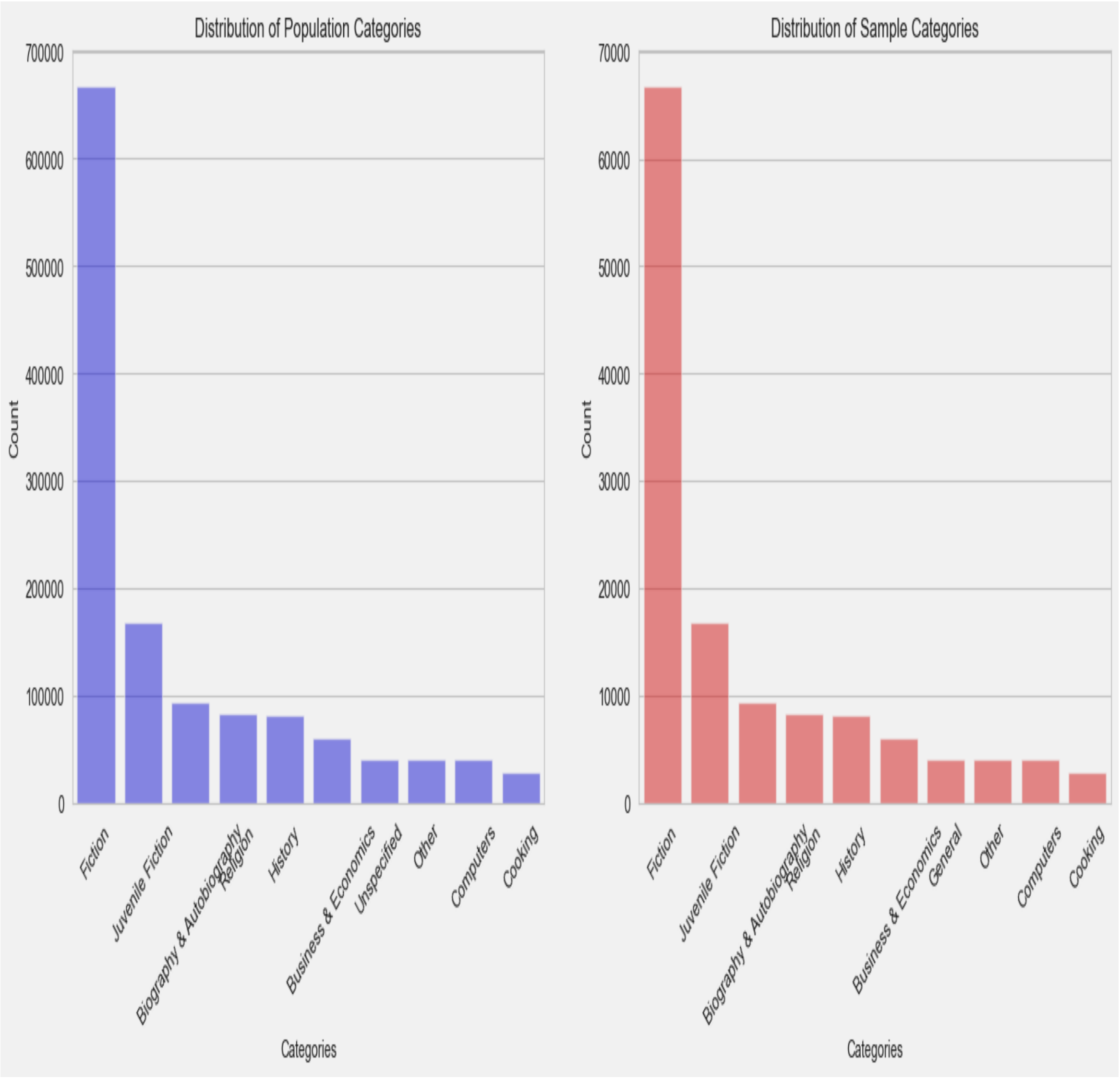
EDA

- The sample used for analysis mirrors the overall population, ensuring it's a valid representation for study.
- A logarithmic scale was used for the 'ratingsCount' to manage the wide range of data better.
- The sample and the population data showed a right skew in 'ratingsCount' and a left skew in 'review/score'.
- The density plots created for these attributes provide a clear visualization of the central trends.
- The sample’s consistency in distribution shapes and skewness confirms a robust sampling methodology which is critical for machine learning accuracy



EDA

- The sample matches the overall category distribution of the full dataset.
- Fiction is the most common category, well-represented in the sample.
- Smaller categories are proportionally included, showing a well-rounded sample.
- The sampling approach ensures an unbiased analysis, even with computational limit



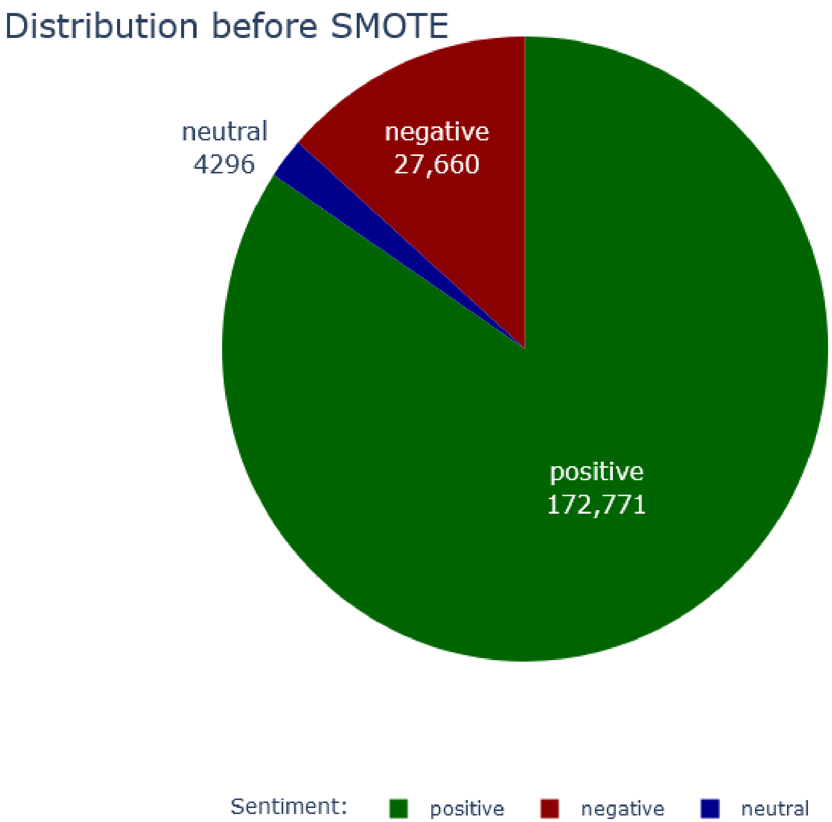
Feature Extraction Techniques

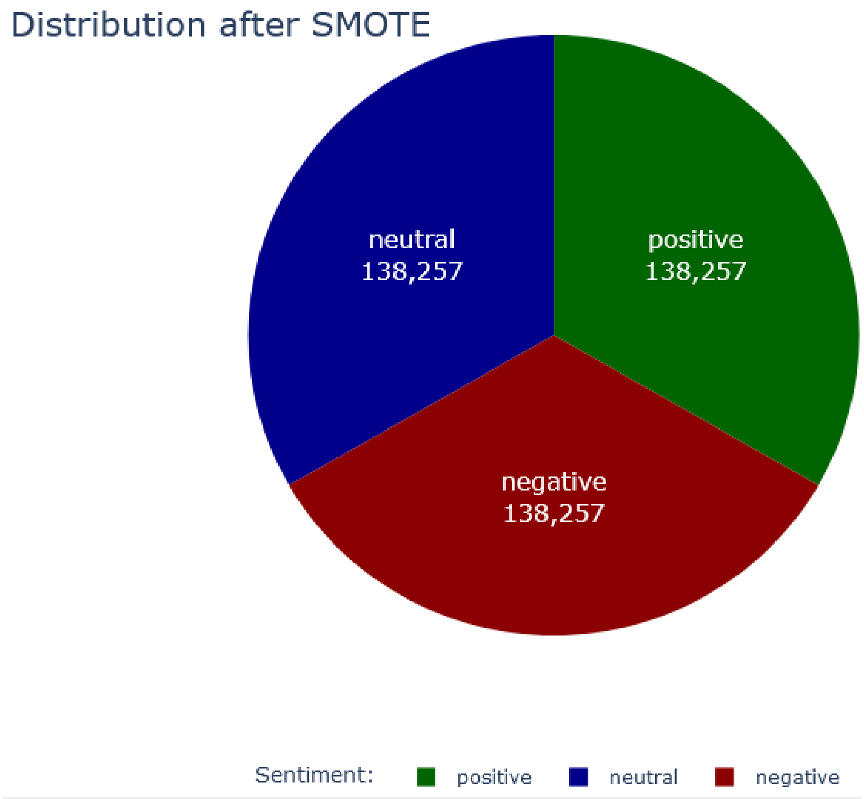
-Feature extraction transforms unstructured text into a structured form that machine learning models can process effectively.

-TF-IDF (Term Frequency-Inverse Document Frequency):
Highlights key terms critical to understanding sentiments within texts, aiding in the differentiation of sentiments by their importance relative to the whole document collection.

-Word2Vec: Goes beyond mere term frequency to capture contextual nuances and semantic relationships between words, thereby refining the sentiment classification process.

Class Imbalance





Ryerson University

Assessing Model Performance with TF-IDF and Word2Vec Features

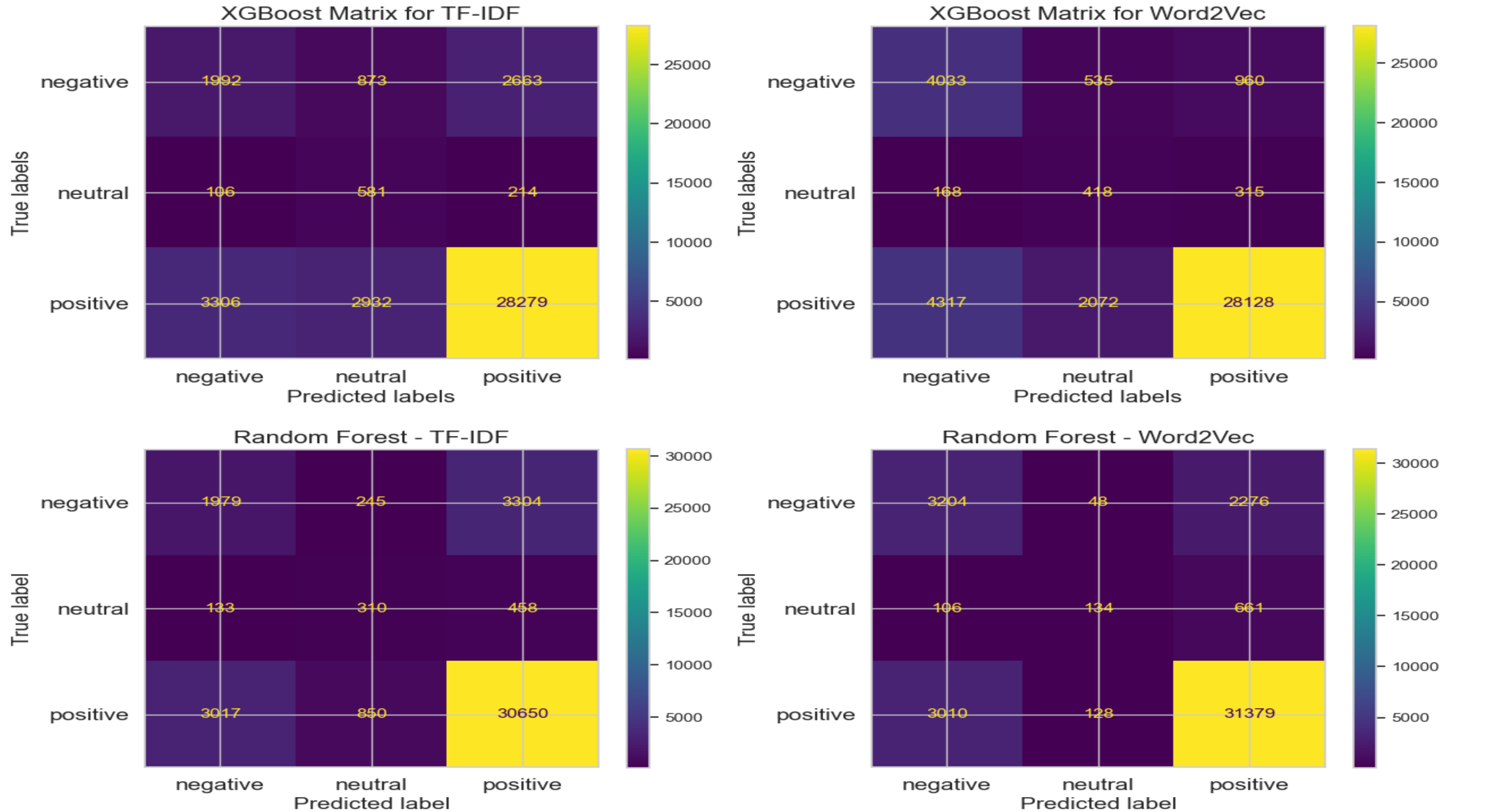
Table 1-Classification algorithms with TF-IDF

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.676476	0.83	0.68	0.73
Naive Bayes	0.59996	0.84	0.60	0.68
Random Forest	**0.804449**	0.81	0.80	0.81
XGBoost	0.75348	0.80	0.75	0.78

Table 2-Classification algorithms with Word2Vec

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.710716	0.88	0.71	0.77
Naive Bayes	0.499316	0.82	0.50	0.58
Random Forest	**0.847872**	0.85	0.85	0.85
XGBoost	0.796317	0.87	0.80	0.82

Model Comparison: Confusion Matrices for TF-IDF and Word2Vec



Confusion matrices reveal that Random Forest and XGBoost models, particularly with Word2Vec features, demonstrate higher accuracy and balanced classification across different sentiment classes.

Ryerson University

Cross Validation

Cross Validation Table 3

Algorithm	Mean Accuracy score	Standard Deviation
Random Forest TF-IDF	92.17%	2.38%
Random Fores Word2Vec	94.94%	0.37%
XGBoost TF-IDF	75.23%	0.35%
XGBoost Word2Vec	89.41%	0.10%

Conclusions

1. **Sentiment Distribution**: "Analysis revealed a predominantly positive sentiment across book reviews, with genre and authorship noticeably influencing sentiment tendencies."
2. **Feature Extraction Success**: "Both TF-IDF and Word2Vec significantly improved sentiment analysis accuracy. Word2Vec, in particular, was instrumental in enhancing model precision."
3. **Model Performance**: "Random Forest and XGBoost models, especially when using Word2Vec features, emerged as top performers in accurately classifying sentiments."

Ryerson University

Limitations & Recommendations

•**Limitations**: Our study faced computational constraints, limiting the depth of analysis. Processing slang and informal language also posed challenges due to the nuances of language use.

•**Recommendations**:

- Future research should leverage high-performance computing resources to expand dataset analysis capabilities."
- Incorporate advanced linguistic models and updated NLP techniques to improve accuracy in processing diverse language expressions.
- Explore sophisticated preprocessing and feature extraction methods, such as BERT or GPT, to further refine sentiment analysis insights.

Ryerson University

Thank You

Ryerson University