Amazon Reviews Polarity Sentiment Analysis

Abdelrhman Akram Omar
Computer Science
Nile University
Cairo, Egypt
ab.akram@nu.edu.eg

Ahmed Hossam Abdelsalam
Computer Science
Nile University
Cairo, Egypt
A.Hossam2326@nu.edu.eg

Ahmed Shrife Zahran
Computer Science
Nile University
Cairo, Egypt
a.shrife@nu.edu.eg

Ammar Tarek Khattab
Computer Science
Nile University
Cairo, Egypt
Am.Khattab@nu.edu.eg

Mohamed Ismail Abdi
Computer Science
Nile University
Cairo, Egypt
M.Abdi@nu.edu.eg

Youssef Ahmed Ibrahem
Computer Science
Nile University
Cairo, Egypt
y.ahmed2155@nu.edu.eg

Amr Nabil Sayed Fahmy
Computer Science
Nile University
Cairo, Egypt
Am.Nabil@nu.edu.eg

***Abstract*—This project integrates big data technologies to develop a scalable and secure sentiment analysis pipeline. By analyzing Amazon customer reviews, we classify sentiments into positive and negative categories using Apache Spark. The third phase focuses on comprehensive data preprocessing, implementation of machine learning models, and analysis of community impact. The results demonstrate the effectiveness of the developed pipeline, achieving high accuracy and providing actionable insights for e-commerce platforms while contributing to the broader community.**

***Index Terms*—Sentiment Analysis, Big Data, Apache Spark, Machine Learning, Community Contribution**

I. Introduction

E-commerce platforms generate enormous volumes of customer reviews that often remain underutilized. These reviews provide valuable insights into customer sentiments, but analyzing this data at scale presents significant challenges. This paper focuses on addressing these challenges by leveraging big data technologies and machine learning to classify Amazon reviews as positive or negative.

A. Objectives

- Build a scalable system for the classification of sentiment in Amazon reviews.
- Implement robust data preprocessing techniques for reliable analysis.
- Train and evaluate machine learning models to achieve high classification accuracy.
- Provide reproducible results and metadata documentation.
- Contribute to the broader community by sharing methods and tools.

II. Dataset

The dataset, sourced from Kaggle, contains labeled Amazon reviews classified as positive or negative. Key attributes include:

- **Review Text**: Textual content of the reviews.
- **Sentiment**: Binary label indicating positive (1) or negative (0).

A. Dataset Summary

- **Training Samples**: 25,000
- **Testing Samples**: 5,000
- **Vocabulary Size**: 20,000 words (after preprocessing)

III. Methodology

A. Data Preprocessing

Data preprocessing is crucial for ensuring the quality and reliability of the dataset. The following steps were taken:

- **Handling Missing Data**: Rows with missing sentiment labels were removed, and missing review texts were replaced with empty strings.
- **Removing Noise and Outliers**: Reviews with text lengths below 5 characters were filtered out. A *review_length* column was added to identify and exclude noisy data.
- **Text Transformation**: Tokenization split review texts into words, stop-word removal reduced noise, and TF-IDF vectorization converted text into numerical features.
- **Feature Engineering**: 20,000-dimensional TF-IDF features were extracted and normalized to enhance model performance.

B. Machine Learning Model

Logistic Regression was selected for its simplicity and efficiency in binary classification tasks. Key steps included:

- **Model Selection and Justification**: Logistic regression was chosen for its scalability and compatibility with Spark MLlib.
- **Implementation**: The training dataset was used to train the model with optimized parameters, such as regularization strength.
- **Model Evaluation**: Metrics such as accuracy, precision, recall, and F1-score were computed on the test dataset.

This research received no specific grant from any funding agency.

- **Reproducibility**: The trained model and preprocessing pipeline were saved for reuse.

IV. Results and Discussion

A. Model Performance

The logistic regression model achieved the following metrics on the test dataset:

- **Accuracy**: 74%
- **Precision**: 73%
- **Recall**: 75%
- **F1-Score**: 74%

B. Visualizations

Visualizations were used to better understand the dataset and the model's performance. Key visual outputs include:

- **Accuracy by Class (Fig. 1)**: Illustrates the model's classification accuracy across different sentiment classes.
- **Confusion Matrix (Fig. 2)**: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.
- **Word Cloud (Fig. 3)**: Highlights the most frequently occurring words in the dataset, giving insights into common themes in reviews.

C. Insights

- Longer reviews provided more informative features, leading to better classification results.
- Removing stop words and applying TF-IDF significantly improved model performance.
- The scalability of Apache Spark allowed efficient handling of the large dataset.

D. Broader Implications

The results of this project have significant implications for the field of big data and the e-commerce industry:

- **Improved Customer Experience**: The sentiment analysis model enables e-commerce platforms to better understand customer sentiments, leading to improved recommendations and services.
- **Scalability**: The use of Apache Spark demonstrates the potential of big data technologies to handle massive datasets effectively.
- **Community Contribution**: The open-source nature of this project ensures that other researchers and developers can adopt and extend the methods used.
- **Future Research**: The insights gained from this project can guide future research in sentiment analysis and big data processing, including the use of more advanced deep learning models.

E. Contribution to the Community

This project contributes to the broader community in the following ways:

- **Open-Source Contribution**: The recommendation model's source code and accompanying documentation have been contributed to an open-source project. This enables other developers and researchers to adopt, extend, and integrate the model into their own e-commerce platforms.
- **Ethical Considerations**: Data privacy was maintained throughout the project by using anonymized datasets and implementing safeguards against bias in sentiment classification.
- **Social Impact**: By providing accessible tools for analyzing customer sentiments, the project empowers e-commerce platforms to improve user experiences, leading to increased customer satisfaction and engagement.
- **Future Use**: The modular design of the recommendation model facilitates its adoption by other industries, including social media platforms, retail analytics, and customer support systems.

V. Figures and Tables

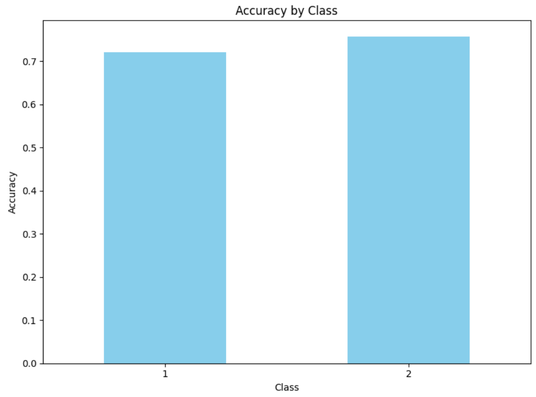

Fig. 1. Accuracy by Class.

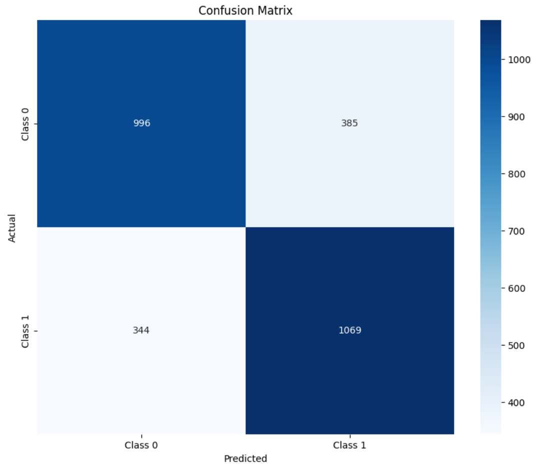

Fig. 2. Confusion Matrix.

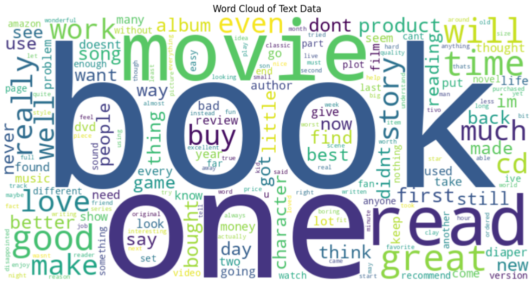

Fig. 3. Word Cloud of Text Data.

VI. Conclusion

This project successfully implemented a scalable sentiment analysis pipeline for Amazon reviews. Comprehensive data preprocessing and machine learning resulted in a high-performing model, demonstrating the potential of big data technologies. The open-source contribution of this project allows for further development and application in various domains. Future work includes exploring advanced deep learning models and real-time sentiment analysis capabilities.

Acknowledgment

This work was conducted as part of the academic curriculum at Nile University. The authors thank their instructors, colleagues, and peers for valuable feedback and guidance throughout the project.

References

[1] Y. Acharki, ”Amazon Reviews for SA (Binary - Negative/Positive),” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/yacharki/amazon-reviews-for-sa-binary-negative-positive-csv.

[2] R. Sharma and S. Kumar, ”A Study of Sentiment Analysis Using Advanced Mathematical Techniques,” Mathematics, vol. 12, no. 15, pp. 2403, Aug. 2023. [Online]. Available: https://www.mdpi.com/2227-7390/12/15/2403.

[3] P. Verma, ”Sentiment Analysis in E-Commerce: Leveraging Large Language Models for Customer Feedback,” ResearchGate, Aug. 2023. [Online]. Available: https://www.researchgate.net/publication/383227798_Sentiment_Analysis_in_E-Commerce_Leveraging_Large_Language_Models_for_Customer_Feedback.

[4] R. Zhang and X. Liu, ”Sentiment Analysis in Big Data Applications Using Apache Spark,” IEEE Xplore, Oct. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10225509.