International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Data Exploration and Modeling using Machine-Learning

Guide: Dr. V Shiva Nagaraju, Professor, ECE & IARE

V.Rishitha[1], B.Neelima[2]

[1]V.Rishitha, ECE & IARE
[2]B.Neelima, ECE & IARE

Abstract - Data exploration and modelling of Amazon product reviews offers valuable insights into consumer opinions and satisfaction levels, enabling businesses to improve products and services. This study employs natural language processing (NLP) techniques to analyze a large dataset of Amazon reviews, focusing on sentiment polarity (positive, negative, or neutral) and opinion trends. Using advanced machine learning models such as Support Vector Machines (SVM), Naïve Bayes, and deep learning algorithms like LSTM (Long Short-Term Memory), the research achieves high accuracy in sentiment classification. Preprocessing steps, including tokenization, stop-word removal, and stemming, ensure the data is optimized for analysis.

The findings reveal key patterns in customer sentiment, highlighting factors that drive satisfaction or dissatisfaction. Products with high ratings consistently exhibit reviews rich in positive sentiments, emphasizing quality, affordability, and customer service. Conversely, negative sentiments are predominantly associated with issues like delayed shipping, poor durability, or misleading product descriptions. The study also demonstrates how sentiment analysis can identify subtle differences in consumer feedback across product categories, helping businesses tailor strategies to specific markets.

***Key Words*:** data exploration, data modelling, feedback, percentage, templates, journals

INTRODUCTION

Data exploration and modelling are essential steps in the data science lifecycle, bridging the gap between raw data and actionable insights. This project focuses on leveraging data to uncover patterns, relationships, and insights, and subsequently building models to predict, classify, or describe outcomes .

The data exploration phase is the foundation of this work. It involves understanding the dataset's structure, identifying trends, outliers, and missing values, and determining the relationships between features. Key tools and techniques, such as statistical summaries, data visualizations, and correlation analyses, are applied to make the data ready for modelling.

The modelling phase builds upon the insights gained during exploration. It includes selecting appropriate algorithms, training models, and validating their performance. The project emphasizes the importance of an iterative approach: refining models based on evaluation metrics to ensure accuracy, interpretability, and robustness.

Data analysis on amazon product reviews has emerged as a powerful tool for understanding customer feedback in various industries, and e-commerce is no exception. Amazon, being one of the largest online retailers, accumulates millions of product reviews from consumers daily. These reviews contain valuable insights into customer satisfaction, product performance, and overall market trends. Analyzing the sentiment expressed in these reviews can help companies better understand consumer needs, improve product quality, and enhance their marketing strategies.

Data analysis on amazon product reviews involves using natural language processing (NLP) and machine learning techniques to identify and quantify emotions expressed in textual data. By categorizing reviews as

positive, negative, businesses can gauge public perception and make data-driven decisions. For instance, a product with predominantly positive reviews may be indicative of high customer satisfaction, while a surge in negative feedback could signal potential quality issues or unmet expectations.

Tools and Technologies Used

Tools Used:

1. **Text Preprocessing Tools:** Libraries like NLTK, spaCy, and TextBlob are used to clean and prepare text. This includes tokenization, stop-word removal, lemmatization, and stemming.
2. **Feature Extraction:** Tools like TF-IDF Vectorizer or CountVectorizer in scikit-learn help convert textual data into numerical representations. Advanced embedding techniques like Word2Vec, GloVe, or BERT provide contextual understanding of words.
3. **Sentiment Lexicons:** Prebuilt dictionaries like VADER, SentiWordNet, or TextBlob can determine the sentiment polarity (positive, negative, neutral) of words or phrases.
4. **Machine Learning Frameworks:** Libraries like scikit-learn, TensorFlow, PyTorch, or Keras are **used** to build and train classification models such as Logistic Regression, SVMs, or deep learning networks (e.g., LSTMs or Transformers).
5. **Visualization Tools:** Libraries such as Matplotlib, Seaborn, or Plotly help visualize sentiment trends.
6. **Cloud Services:** Platforms like AWS Comprehend or Google Cloud NLP provide ready-to-use APIs for sentiment analysis**.**

Technologies Used:

Numpy - Supports large, multi-dimensional arrays and numerical operations.

Matplotlib - used for creating static, interactive, visual plots.

Seaborn - Python library for creating informative, visually appealing statistical plots and data visualizations**.**

scikit-learn - A robust library for implementing machine learning algorithms.

Wordcloud - word cloud visually represents text data, highlighting frequently occurring words to identify themes and trends.

Nltk- text analysis, NLP tasks like tokenization, parsing, and sentiment analysis**.**

Xgboost - Building efficient, scalable machine learning models, particularly for regression and classification tasks.

Streamlit- Simplifies the creation of interactive machine learning web applications.

Flask - Building lightweight, scalable web applications and APIs with Python.

Table -1: Sample Table format

	rating	date	variation	Verified reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1
2	4	31-Jul-18	Walnut Finish	"Sometimes while playing a game, you can answer...	1
3	5	31-Jul-18	Charcoal Fabric	"I have had a lot of fun with this thing. My 4...	1
4	5	31-Jul-18	Charcoal Fabric	Music	1

Methodology

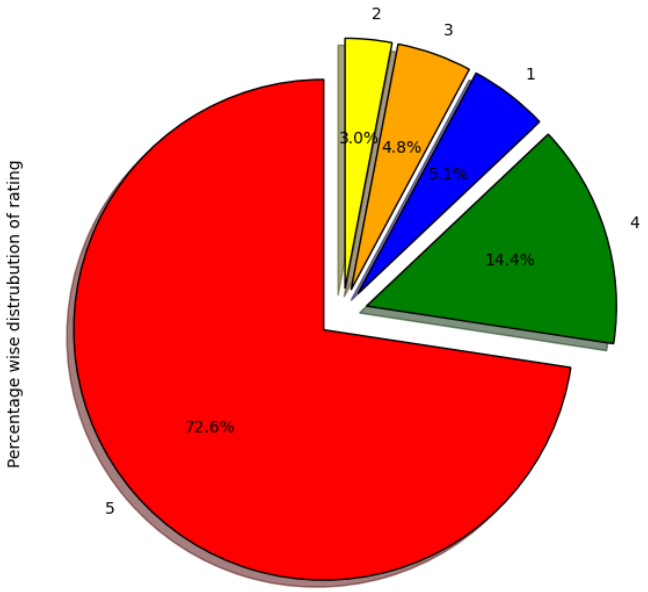
The methodology for Data analysis And modelling using ai on Amazon product reviews involves key steps:

1. **Data Collection**: Scrape or download product reviews from Amazon using APIs or web scraping tools. The data typically includes review text, ratings, and other metadata.
2. **Preprocessing**: Clean the data by removing HTML tags, special characters, stopwords, and converting text to lowercase. Tokenization,

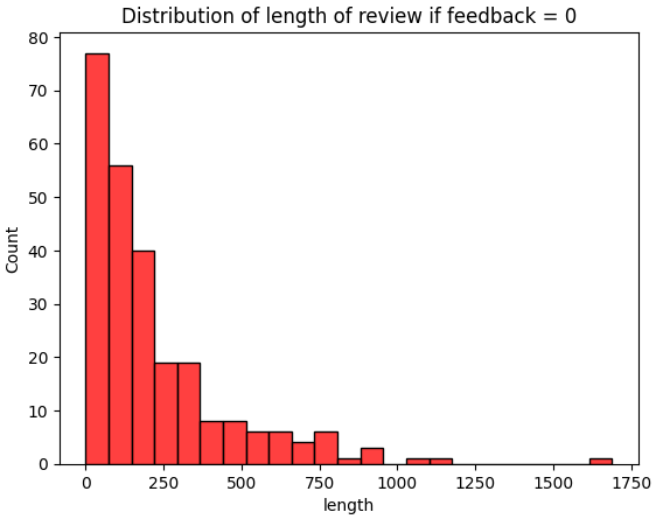
lemmatization, or stemming may also be applied to standardize the text.

3. **Labeling Sentiments**: Assign sentiment labels (e.g., positive or negative) based on review ratings (e.g., 4-5 stars = positive, 1-2-3 stars = negative) or manual annotation if needed.
4. **Feature Extraction**: Convert text data into numerical representations using techniques like Bag-of-Words, TF-IDF, or word embeddings
5. **Model Selection**: Train machine learning models such as Logistic Regression, SVM, Naive Bayes, or deep learning models on the labeled dataset.
6. **Evaluation**: Assess model performance using metrics like accuracy, precision, recall, and F1-score on a test dataset.
7. **Deployment**: Integrate the trained model into applications for real-time data analysis and modelling using ai of Amazon reviews.

Fig -1: Figure



Charts



CONCLUSIONS

In conclusion, the completion of a data exploration and modeling project in machine learning marks the culmination of a structured process aimed at deriving insights and creating predictive models from raw data. This project demonstrates the importance of a systematic approach to data analysis and machine learning model development. The outcomes of this project not only provide actionable insights but also offer a scalable and reliable solution for [specific application, e.g., predicting customer churn, classifying medical conditions, optimizing resource allocation]. Moreover, the model's explainability was enhanced through tools like SHAP or LIME, making it more interpretable for stakeholders.

ACKNOWLEDGEMENT

This research and the development of the data analysis and modelling system would not have been possible without the invaluable contributions of numerous individuals and organizations. We would like to express our deepest gratitude to everyone who supported us throughout the completion of this project on *"Data Analysis And Modelling Using Machine Learning ."*First and foremost, we extend our sincere thanks to [**Dr. V Shiva Nagaraju**] for their invaluable guidance, insightful suggestions, and continuous encouragement, which significantly contributed to the success of this project.

We are also immensely grateful to [Institute of Aeronautical Engineering] for providing the necessary resources, tools, and a conducive environment to carry out this research.

Additionally, we acknowledge the contribution of publicly available datasets, APIs, and tools which were pivotal in the analysis and implementation of this project.

.**REFERENCES**

1. Bing Liu, "Exploring User Opinions in Recommender Systems", *Proceeding of the second KDD workshop on Large Scale Recommender System and the Netflix Prize Competition*, April 2012.
2. Antonio Moreno-Ortiz and Javier Fernández-Cruz, "Identifying polarity in financial texts for sentiment analysis: a corpus-based approach", *7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)*.
3. Zhang Wenhao, Hua Xu and Wan Wei, "Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis", *Expert Syst Appl*, 2012.
4. Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 417-424, July 2002.
5. Bo Pang and Lillian Lee, "Seeing stars Exploiting class relationships for sentiment categorization with respect to rating scales", *Proceedings of the ACL*, 2005.
6. Theresa Wilson, Janyce Wiebe and Paul Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", *Advanced Research and Development Activity (ARDA)*.
7. State university Xing Fang and Justin Zhan, "Sentiment Analysis using product review data" in Journal of Big data, North Carolina A T, Greensboro, NC, USA: Springer, 2015.
8. Subha brata Mukherjee and Pushpak Bhattacharyya, "Feature Specific Sentiment Analysis for product Reviews", *IET 2015 IIT Bombay*.
9. Lakka raju, Chiranjib Bhattacharyya, Indrajit Bhattacharyya and Srujana Merugu, "Exploiting (ICICCT 2017) Coherence for the simultaneous discovery of latent facts and associated sentiments", *SIAM International Conference on Data Mining (SDM)*, April 2011.
10. Min qing Hu and Bing Liu, "Miming and Summarizing customer reviews", *KDD 04: proceedings of the tenth ACM SIGKDD international Conference on knowledge discovery and data mining*.