Regional Product Sentiment Analysis Using Clustering, Classification, Statistics and NLP

Abstract
This abstract provides an analysis of customer sentiment towards iPhone products based on reviews collected across multiple continents, including APAC, EMEA, and the United States. Using a comprehensive dataset, the relationships between various product features (e.g., variant, verification status, and country/continent) and customer ratings are explored. The study leverages machine learning models, traditional classification methods, and modern large language models (LLMs) to assess sentiment and identify key factors contributing to customer satisfaction or dissatisfaction (Devlin et al., 2018; Le & Mikolov, 2014).

In addition to conventional sentiment analysis, clustering techniques are used to explore how customer sentiments differ across geographic regions. Country-based sentiment clustering aims to uncover distinct user preferences, satisfaction levels, and trends that may vary by country, providing valuable insights into how geographic location affects product perception.

Hypothesis testing is applied to assess the statistical significance of differences in sentiment based on various product attributes, enabling a deeper understanding of user behavior. By combining predictive modeling, hypothesis testing, and clustering, the solution offers a detailed examination of customer preferences, supporting data-driven decision-making for targeted improvements and marketing strategies (Mikolov et al., 2013; Pennington et al., 2014).

Problem Statement
The purpose of this analysis is to understand customer sentiment regarding iPhone products and identify the major factors influencing customer satisfaction or dissatisfaction across different regions and product variants. The specific objectives of this project are as follows:

1. **Understand Sentiment Variation Across Countries or Continents**:
    - Perform **country-based sentiment analysis using clustering** to determine how customer satisfaction levels differ by region.
    - Identify **country-specific clusters** of customers with similar sentiments to understand the unique needs and preferences in each geographic area.
2. **Identify Key Factors Influencing Ratings**:
    - Analyze relationships between different product features (e.g., color, size, service provider) and customer ratings to identify attributes that have the greatest influence on sentiment.
    - Conduct **hypothesis testing** to validate the statistical significance of observed differences in sentiment based on country, product variant, and verification status (Le & Mikolov, 2014).
3. **Compare Traditional Machine Learning Models with LLMs for Sentiment Prediction**:

    - Compare the efficacy of traditional machine learning models (e.g., Random Forest, Logistic Regression) with fine-tuned **large language models** (e.g., DistilBERT) for predicting customer sentiment.
    - Evaluate each model's ability to interpret complex language, understand context, and classify sentiment (Collobert et al., 2011).
4. **Temporal Trends Analysis**:
    - Investigate temporal trends to understand how sentiment and ratings evolve over time (e.g., during product release cycles or seasonal effects).

Research Questions to be answered in this analysis:

- Are there significant differences in customer sentiment across different countries?
- What are the primary drivers of customer satisfaction or dissatisfaction in different regions?
- How do customer sentiment trends change over time?
- Can advanced language models outperform traditional machine learning models in predicting sentiment?

Clustering for Country-Based Sentiment Analysis
To address research questions related to geographic differences in customer sentiment, clustering techniques will be employed. By applying **unsupervised clustering** methods such as **K-Means** or **Hierarchical Clustering** to customer reviews, the study aims to identify groups of countries that exhibit similar sentiment patterns toward the product (Lloyd, 1982). By forming clusters based on sentiment, the analysis will:

- Identify geographic regions where customers report high or low satisfaction levels.
- Understand regional preferences, which could be influenced by factors like pricing, cultural differences, or consumer expectations.

This clustering approach, combined with sentiment analysis, will support the development of targeted marketing strategies, product customization, and feature improvements that cater to regional preferences.

Dataset Overview

- **Total Entries**: 3,062
- **Number of Columns**: 11

Features:

1. **productAsin**: Product identifier for the iPhone variant.
2. **country**: Country where the review was written.
3. **date**: Date of the review.
4. **isVerified**: Indicates whether the review is verified.

5. **ratingScore**: Rating given by the reviewer (integer values).
6. **reviewTitle**: Title of the review.
7. **reviewDescription**: Detailed description of the review (with some missing values).
8. **reviewUrl**: URL of the review (with some missing values).
9. **reviewedIn**: Details on where and when the review was conducted.
10. **variant**: Information about the product variant, including color and size.
11. **variantAsin**: Identifier for the product variant.

Algorithms to be utilized for answering key research questions

Clustering: K-Means, K-Means++

Classification: Logistic Regression, Decision Tree, Random Forest

NLP: **Distilber**

Analysis Overview:

1. **Preprocessing (Dataset Cleaning)** :
    - Perform data cleaning, handling missing values, and normalization to prepare the dataset for analysis.
2. **Exploratory Data Analysis (EDA)**:
    - Perform clustering and **hypothesis testing** to gain insight into the data.
    - Use **K-Means clustering** to evaluate and compare performance (e.g., KNN, K-Means++, etc.).
3. **Storage Size-Based Analysis (Clustering)**:
    - Analyze different storage sizes (e.g., 128GB, 256GB) for their popularity across various regions.
4. **Classification**:
    - Use **Logistic Regression**, **Decision Tree**, and **Random Forest** models for classification tasks (Phase 1).
5. **Natural Language Processing (NLP)**:
    - Utilize models such as **BERT** for advanced sentiment classification.
    - Perform in-depth analysis at different time intervals to understand changes in sentiment and rating.
6. **Temporal Analysis**:
    - **One Month After Release**: Analyze sentiment immediately after release to determine skew and limitations of the dataset.
    - **Two Months After Release**: Evaluate changes in sentiment.
    - **One Year After Release**: Perform a detailed analysis, including metrics such as confusion matrices and heatmaps.
7. **Modeling Comparison**:

- Compare different models for their efficiency in sentiment prediction, accuracy, and ability to handle context effectively.

Repository links:

https://github.com/gshiva1975/AAI-501/edit/main/README.md

Dataset Location:

https://www.kaggle.com/datasets/mrmars1010/iphone-customer-reviews-nlp

Team Contribution Table

Tasks	**Team Member1** **Gangadhar Singh Shiva**	**Team Member2** **Akshobhya Rao BV**
Create and Access Github Account	**yes**	**yes**
Investigate/Brainstorming on Dataset	**yes**	**yes**
Data Cleansing	**yes**	**yes**
K Means, K Means++ Clustering for regional grouping	**yes**	**yes**
Hypothesis Testing for regional grouping	**yes**	**yes**
ML Classification Analysis - Decision Tree, Random Forest , Ada Boost Etc Sentiment Analysis	**yes**	**yes**
BERT NLP Sentiment Analysis	**yes**	**yes**

Comparison,Technical Report/Presentation	**yes**	**yes**

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805. https://arxiv.org/pdf/1810.04805
2. Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*. International Conference on Machine Learning. https://cs.stanford.edu/~quocle/paragraph_vector.pdf
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
4. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.1162/tacl_a_00349
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). *Natural Language Processing (Almost) from Scratch*. Journal of Machine Learning Research, 12, 2493-2537. https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf
6. Lloyd, S. (1982). *Least Squares Quantization in PCM*. IEEE Transactions on Information Theory, 28(2), 129-137. https://ieeexplore.ieee.org/document/5453745