Amazon Reviews Summarizer

Team Members: M. Dheeraj - 2101AI18
R. Vivek - 2101CS65
R. Eshwar - 2101AI25

Problem Statement

- Identify whether a product is useful in India based on user reviews.

1. Collect 500 reviews from Indian users for each of 50 products and annotate usefulness.
2. Analyze sentiment and mine product aspects (likes/dislikes).
3. Summarize user opinions and train a classifier.

Dataset Details

1. A review scraper using selenium and beautifulsoup to get reviews with features product_id , user_name , review_rating , review_title , review_description and save them to data.csv

2. Dataset size : 25000 (50 * 500) (Products * Reviews per product)

Annotation

- Performed Sentiment Analysis using:
  - **TextBlob** – Lexicon-based polarity , **VADER** – Rule-based compound score
  - **BERT** – Transformer-based sentiment , **Rating (scaled)**
- **Final Sentiment Score =** Average of all methods
- Products labeled as:
  - **"Useful"** if average score > 0.705
  - **"Not Useful"** otherwise

Reviews Analysis

- **Sentiment Analysis**:

  Classified reviews as *Positive*, *Negative*, or *Neutral* using **VADER**.

  Overall Verdict: Products labeled *Liked* if positive reviews > negative, else *Not Liked*.

- **Aspect-Based Opinion Mining**:

  Used **spaCy** to extract key product aspects (e.g., *battery*, *camera*).

  Counted sentiment mentions for each aspect.

  Listed top appreciated and criticized aspects for each product.

Methodology (1)

- **Data Preprocessing:**
  - Merged datasets and engineered features (e.g., average ratings, sentiment scores).
  - Created target variable: "Useful" vs "Not Useful."
- **EDA:**
  - Visualized **Verdict Distribution**, **Correlation Heatmap**, **Sentiment Score Distribution**, and **Review Length Distribution**.

Methodology (2)

Modeling:

- Trained **Random Forest** classifier.
- Preprocessed data using **TF-IDF** for text and **StandardScaler** for numeric features.

Evaluation:

- Assessed performance with **Confusion Matrix**, **ROC Curve**, and metrics (Accuracy, Precision, Recall, F1-Score.

Results

- **Annotation :**
  - Useful 41
  - Not Useful 9
- Sentiment Analysis and Aspect-based opinion mining :
  - Generated two csv files to report the findings

  - **Product_overall_sentiment_summary.csv** includes the following columns:
  - product_id, total_reviews, positive_reviews, negative_reviews, neutral_reviews, overall_verdict, top_appreciated_aspects, top_criticized_aspects.

  - **Aspect_based_opinion_per_product.csv** Includes the following columns:
  - product_id, aspect, positive_mentions, negative_mentions, net_sentiment.

Results (2)

Model	Accuracy	Precision (Useful)	Recall (Useful)	F1-score (Useful)
Random Forest	0.9497	0.9710	0.9668	0.9689

OUTPUT \ TARGET	Class0	Class1
Class0	1743 16.646%	282 2.693%
Class1	245 2.340%	8201 78.321%

Conclusion

Future Work Scope :

- Gather more diverse data from a wider range of products, which will help improve the generalizability of the model.
- Enhance the aspect extraction process to cover more detailed aspects.
- Use model’s output in recommender systems for Indian zone as a feature.

Contributions :

- R.Eshwar (2101AI25) - Review Scraping , Automated annotations.
- M.Dheeraj (2101AI18) - Sentimental Analysis and Aspect-based opinion mining.
- R.Vivek (2101CS65) - Train Classifier based on dataset and extracted features.
- https://github.com/eshwar0210/CS563-NLP

Thank you!

Team Members: M. Dheeraj - 2101AI18
R. Vivek - 2101CS65
R. Eshwar - 2101AI25