Dataset Report

Dataset Name – Amazon Fine Food Reviews

Source – Kaggle: https://www.kaggle.com/snap/amazon-fine-food-reviews/downloads/amazon-fine-food-reviews.zip

Description: Reviews for fine foods sold on Amazon, with ratings from 1–5 stars.

1.Dataset Overview:

This dataset is the Amazon Fine Food Reviews dataset. It contains reviews of fine foods from Amazon. The data includes review metadata such as reviewer information, product details, review text, and ratings.

Imbalanced Shape (568454, 14)

2.Dataset Shape:

Imbalanced Shape: (568454, 14)

3.Column Details:

-Id – Unique identifier for each review

- ProductId – Unique identifier for the product
- UserId – Unique identifier for the user
- ProfileName – Name of the reviewer
- HelpfulnessNumerator – Number of users who found the review helpful
- HelpfulnessDenominator – Number of users who indicated helpfulness
- Score – Rating given by the user (1–5 scale)
- Time – Timestamp of the review (Unix time)
- Summary – Short summary of the review
- Text – Full review text
- Rating – Derived label column for training
- Review_text – Processed review text
- ReviewLength – Length of the review text

4. 📊 Rating Distribution

- ⭐ 1 = 36,275
- ⭐ 2 = 20,791
- ⭐ 3 = 29,754
- ⭐ 4 = 56,041
- ⭐ 5 = 250,714

5. Data Characteristics

- **Average review length:** 3 words
- **Maximum review length:** 3432 words
- **Missing data:** NO
- **Language/Encoding:** Mostly English text.

6,⭐ Most Common and Least Common Words per Rating

Rating	Most Common Words	Least Common Words
1★	bad, disappointed, awful, terrible, waste, poor, horrible, not, expensive, disgusting	love, delicious, amazing, excellent, perfect, great, good, best, tasty, wonderful
2★	okay, poor, bland, disappointed, not, cheap, average, boring, waste, packaging	love, amazing, delicious, excellent, perfect, great, good, tasty, wonderful, best
3★	average, okay, fine, decent, not, price, flavor, quality, delivery, expect	terrible, horrible, awful, disgusting, waste, bad, not, poor, bland, disappointing
4★	good, nice, quality, taste, love, flavor, product, satisfied, fresh, well	terrible, horrible, awful, disgusting, waste, bad, cheap, poor, boring, disappointing
5★	love, great, excellent, perfect, delicious, amazing, best, good, wonderful, tasty	bad, horrible, awful, disgusting, waste, poor, boring, cheap, disappointing, not

7.Most used words in the dataset

	Word	Count
0	br	181797
1	like	170334
2	good	134979
3	taste	116106
4	one	115513
5	great	112324
6	product	106019
7	flavor	96549
8	coffee	95391
9	tea	90133
10	would	84130
11	love	83441
12	get	73562
13	really	68858
14	food	66286
15	dont	64946
16	much	63780
17	use	61541
18	also	59107
19	little	57302
20	time	56183

8. Data Preprocessing

- Lowercasing all text.
- Removing punctuation, numbers, special characters.
- Removing stopwords.
- Tokenization and padding for deep learning.
- Combining Summary and Text if needed.
- Encoding labels (1–5 stars).

9. Challenges / Limitations

- Imbalanced dataset → more positive reviews than negative.
- Some reviews are very short → limited sentiment info.
- Spelling mistakes, slang, or emojis may affect NLP model performance.
- Reviews may contain neutral language, making 3★ prediction tricky.

10. Future Enhancements

- Use advanced NLP models (BERT, RoBERTa) for better understanding.
- Aspect-based sentiment analysis (taste, packaging, delivery).
- Multi-language support.
- Remove spam or bot reviews for cleaner data.