MILESTONE 1

PROJECT OVERVIEW

•**Project Name:** Review Sense – Extracting Insights from Customer Feedback.
•**Objective:** To build an automated system that cleans real-world, messy customer reviews to prepare them for Sentiment Analysis and Topic Modeling.
•**Dataset Source:** Amazon Customer Reviews (fast Text format)

THE PROBLEM (RAW DATA CHALLENGES)

- **Noise:** URLs, HTML tags, and special characters.
- **Redundancy:** Numbers and punctuation that don't add sentiment value.
- **Stop words:** Common words (is, the, and) that clutter the analysis.
- **Inconsistency:** Mixed casing (Lower vs. Upper).

THE PREPROCESSING PIPELINE

Normalization: Converting all text to lowercase.

Noise Removal: Using Regular Expressions (Regex) to strip URLs and digits.

Tokenization & Filtering: Splitting text into words and removing the STOPWORDS list.

TECHNICAL STACK

- **Language:** Python.
- **Libraries:** pandas: For structured data manipulation.
- re: For complex pattern matching (Regex).
- bz2 & openpyxl: For handling compressed source files and Excel formats.

RESULTS & VERIFICATION

Feature	Raw Feedback (Input)	Cleaned Feedback (Output)	Verification Result
Lowercasing	Stuning...	stuning...	✅ Success
Punctuation	Amazing**!:** This...	amazing...	✅ Removed !:
Stopwords	...**the** non-gamer**...**	...nongamer ...	✅ Removed the
Noise	This sound tra...	sound track...	✅ Normalized spacing

THANK YOU