**AI ML Internship Log**

# Day 3 - Code Understanding and Revision Notes

# Date - 11 June 2025

# Team Role - Member

# Project Title - Personality Prediction from Social Media

# **Goal of the Code**

# To clean raw text by removing noise (like punctuation and stopwords) and extract **meaningful words (tokens)** that can be used for machine learning models — specifically for our project on **personality prediction from social media**.

# **What is Text Preprocessing?**

# Text preprocessing is the process of **cleaning and preparing raw text** so that it can be understood and used by machine learning models.

# Think of it like **cleaning vegetables before cooking** — raw text is messy, and we must remove noise (punctuation, filler words, etc.) before it's useful.

# **Why is Text Preprocessing Important?**

# ML models **don’t understand language** like we do — they need clean, consistent **numeric input**. Preprocessing helps turn raw, noisy text into clean tokens that can later be **vectorized** (using TF-IDF or embeddings) and fed into a model.

# Raw social media content contains:

## Capital Letters

## Emojis, punctuations, hashtags

## Common unimportant words like is, the, and.

# 

# 🧰 **Libraries Used in the Code**

|  |  |
| --- | --- |
| **Library** | **Purpose Used for** |
| nltk | Main NLP library (downloading stopwords) |
| stopwords (from nltk.corpus) | Provides list of common English words like "is", "and" |
| String | Contains string.punctuation used to remove symbols. |

# **5.** 🔄 **Preprocessing Steps Performed**

# 

|  |  |  |
| --- | --- | --- |
| **Step** | **What It Does** | **Why it’s Needed** |
| 1.Lowercase | Converts all text to lowercase. | Ensure words like Love and love are the same. |
| 2. Remove Punctuation | Deletes punctuation symbols like !, . | These don't carry meaning |
| 3. Tokenization | Splits sentence into words using split() | Words are easier to process individually |
| 4. Remove Stopwords | Removes words like "is", "and", "the" | Focus on meaningful words only. |

# 

# 

# **6.** 💬 **Final Cleaned Output**

# **Original text:**

# "I LOVE AI and Machine Learning!! It's so exciting."

**After preprocessing:**

['love', 'ai', 'machine', 'learning', 'exciting']

# **7. 🧠 Key Learning from Code**

# Even simple preprocessing improves the quality of data for analysis.

# Understanding **why each step is done** is more important than just running the code.

# This cleaned list of tokens is what we will pass to the **feature extraction** step next (like TF-IDF).

# **8. 💭 How I Feel Now**

# After revisiting the code and understanding what each line does, I feel more confident and clear about the purpose of text preprocessing. I no longer feel like I’m blindly running code — I now understand **why** each step matters. This understanding gives me a stronger base to move forward with the next stage of the project.

# I’m also realizing that **errors and confusion are part of learning**, and what matters most is staying consistent, asking questions, and moving forward step by step.

# 

# 

# **9 . ✅ Conclusion**

# Today was a turning point in my project. I started with confusion and doubt, but by breaking down the code and understanding its logic, I’ve built a solid foundation.

# I now know what text preprocessing is, why it’s essential in NLP, and how it transforms raw data into meaningful input for machine learning.

# I may not know everything yet, but I’m no longer stuck — I’m learning with purpose. This progress, however small, is meaningful, and I’m motivated to continue.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

## 

# 