Customer Support Ticket Cleaning & Annotation System

1. Introduction:

Customer support systems generate large volumes of unstructured text data from emails and chat messages. These support tickets often contain spelling mistakes, inconsistent formatting, abbreviations, and irrelevant content. Such noisy data makes automation difficult and reduces the performance of machine learning models.

This project develops a Natural Language Processing (NLP) pipeline to clean, preprocess, and annotate customer support tickets, converting raw text into structured, machine-learning-ready data.

---

2. Problem Statement:

A SaaS company receives thousands of customer support tickets daily. These tickets are:

- Unstructured
- Noisy and inconsistent
- Contain spelling mistakes
- Hard to classify automatically

Without proper preprocessing, automation fails, classification models perform poorly, and manual ticket handling becomes costly and inefficient.

---

3. Objectives:

- Clean noisy customer ticket text
- Correct spelling mistakes
- Perform tokenization and lemmatization
- Remove stopwords
- Extract Named Entities (NER)
- Label tickets into categories
- Create structured datasets for ML tasks

---

4. Technologies Used:

Technology	Purpose
Python	Programming language
Pandas	Data handling
spaCy	NLP processing & NER
NLTK	Tokenization & stopword removal
TextBlob	Spell correction

---

5. System Architecture:

Raw Ticket → Text Cleaning → Spell Correction → Tokenization → Stopword Removal → Lemmatization → NER → Ticket Labeling → Processed Dataset

---

6. Methodology:

Step 1: Data Collection

A dataset of sample customer support tickets is created in CSV format.

Step 2: Text Cleaning

Special characters, punctuation, and extra spaces are removed.

Step 3: Spell Correction

TextBlob corrects common spelling errors.

Step 4: Tokenization

Text is broken into individual words (tokens).

Step 5: Stopword Removal

Common words like *is, the, and* are removed.

Step 6: Lemmatization

Words are reduced to base form (e.g., *running* → *run*).

Step 7: Named Entity Recognition (NER)

Entities like dates, numbers, and organizations are extracted.

Step 8: Ticket Annotation:

Tickets are labeled into categories such as:

- Login Issue
- Billing Issue
- App Issue
- Delivery Issue
- General Query

---

7. Output:

The system generates a structured dataset containing:

- Cleaned text
- Spell-corrected text
- Processed tokens
- Named entities
- Ticket category labels

---

8. Applications:

- Automated ticket routing
- AI chatbots
- Support analytics
- Customer issue classification
- Helpdesk automation

---

9. Advantages:

- Improves data quality
- Reduces manual effort
- Enhances ML model accuracy

- Handles real-world noisy text

---

10. Future Enhancements:

- Add ML classification model
- Build web interface
- Add sentiment analysis
- Real-time ticket prediction

---

11. Conclusion:

This project demonstrates how NLP techniques can transform raw, unstructured customer support tickets into structured and annotated data. High-quality preprocessing improves automation, reduces operational costs, and enhances intelligent support systems.