NLP Analysis of Customer Support Chat Data

1. Preprocessing Methods

Preprocessing is a crucial step to prepare the dataset for NLP analysis. The following steps were applied:

a. Tokenization

- Split the text into smaller units, such as sentences or words.

Example: “My credit card was charged incorrectly. Can you assist with that?” becomes sentences: [“My credit card was charged incorrectly”, “Can you assist with that?”]

Example: "I'm having trouble logging into my account" becomes tokens: ['I', "'m", 'having', 'trouble', 'logging', 'into', 'my', 'account'].

b. Stemming and Lemmatization

- Stemming: Reduces words to their root forms (e.g., "upcoming" → "upcom").
- Lemmatization: Converts words to their dictionary forms while considering context (e.g., " bought" → "buy").

Comparison:

- ✗ Stemming is faster but less precise; it might generate non-words.
- ✗ Lemmatization is more accurate and context-aware, suitable for semantic analysis.

c. Part-of-Speech (POS) Tagging

- Assigns parts of speech to each token to understand grammatical roles.

Example: "logging" as VERB, "account" as NOUN.

d. Named Entity Recognition (NER)

- Extracts meaningful entities like product names, locations, dates, etc.

Example: "Can I return a product bought last week?" identifies:

Entity: "last week" (DATE)

Entity: "product" (OBJECT).

2. Analysis Steps

a. Common Issue Identification

- Analyzed frequent words/phrases using word frequency and n-grams.

Example: Words like "login," "refund," and "defective" frequently appear.

b. Topic Discovery

- Applied basic clustering or keyword extraction to identify themes.
- A TF-IDF Vectorizer identifies the most significant words in each message,
- highlighting keywords that dominate the context.

Example: Topics include billing issues, product returns, and warranty inquiries.

c. Sentiment Analysis

- Classified customer messages as positive, neutral, or negative using:
  - Polarity scores derived from preprocessed text.

Example: "Can I return a defective product?" → Negative.

d. Entity-Based Grouping

- Grouped queries by extracted entities to identify patterns.

Example: Most queries about "refunds" relate to "defective items."

3. Insights

Frequent Concerns:

- Login issues (e.g., "trouble logging in").
- Refund and billing disputes (e.g., "charged incorrectly").
- Product inquiries (e.g., "defective item" and "warranty policy").

Regional/Entity Patterns:

- Certain products (electronics) and dates (holiday season) require more support.

Sentiment Trends:

- Predominantly negative queries indicate dissatisfaction, while neutral queries suggest straightforward inquiries.

4. Recommendations

- Enhance self-service resources (e.g., FAQs for login issues).
- Improve refund and billing processes.
- Tailor chatbot responses for top query types (e.g., defective products, refunds).