Project:

Built a model that can classify customer complaints into predefined categories(10 categories) using Natural Language Processing(NLP) and Neural Networks.

Steps:

Data Preparation

- **Text Processing:** Removed punctuations from the text, tokenized the sentences into words, removed stop words, performed Lemmatization.
- Removed stop words and added 5% character-level noise so that model generalizes more.
- Used Word2Vec (vector size: 100) to get word embeddings
- Engineered features: Took text length and punctuation percentage so that the models can know for which type of complaints the length and percentage of punctuations are more.

Models Developed:

1. **Feed-Forward Neural Network (FFNN)**
   Simple linear classifier using embeddings + engineered features
2. **LSTM Classifier**
   Used Bidirectional LSTM classifier with dropout and two fully-connected layers.

Both models trained using cross-entropy loss with Adam optimizer.

Key Insights:

- Both models achieved over 85% - 90% accuracy**.**
- LSTM outperformed Feed Forward Network by getting better word order and context.
- Data noise had improved generalization.

Challenges Faced:

- **Handling different complaint lengths:** Since complaints varied a lot in size, I used fixed length padding to keep the input consistent across the project.
- **Handling overfitting:** Even the simplest models were overfitting at first. I improved this by adding more data and injecting small amounts of noise to help the model generalize better.
- **matrix shapes**: I kept track of the shapes of all inputs and layers (A simple feedforward network expects 2D input data with shape (batch_size, input_features), whereas an LSTM expects 3D input data with shape (batch_size, sequence_length, input_features) to handle sequential data right).

Dataset changes:

The original dataset caused the model to overfit pretty quickly it didn't have more unique complaints , and the model memorized rather than learning. So, I decided to switch to a better dataset that included actual complaints along with their correct categories. Here I produced large number of different complaints using Large Language Models through the website. This gave the model more realistic examples to learn from and helped it perform better on unseen data.