Project 2 – Proposal

Tejashri Bhilare

MS in Data Science, Bellevue University

DSC680 – Applied Data Science

Professor - Amirfarrokh Iranitalab

April, 2025

Next Word Prediction – Language Model

Topic:

Language modeling focuses on predicting the next word in a sequence, given a preceding context. It plays a fundamental role in natural language processing applications, such as resolving customer inquiries via chat or email.

Business Problem:

In customer service settings—particularly within chat and email support—representatives often face delays when responding to inquiries due to limited domain knowledge. Timely responses are crucial for delivering excellent service and enhancing customer satisfaction.

To address this challenge, business stakeholders aim to develop a model that leverages historical chat and email resolution data to predict the next word as a customer service representative types. This predictive capability would assist representatives in formulating responses more efficiently.

Objective:

Develop a next-word prediction model based on prior conversational context to assist customer support representatives in real-time communications.

Dataset:

The project will utilize the *Relational Strategies in Customer Service (RSiCS)* dataset, which includes human-computer interaction logs from three live customer service Intelligent Virtual Agents (IVAs) in the domains of travel and telecommunications. Additionally, it incorporates user interactions from the Airline forums on TripAdvisor (August 2016). The dataset sources include:

- TripAdvisor Airline Forum
- Train Travel IVA
- Airline Travel IVA
- Telecommunications Support IVA

This dataset is sourced from Kaggle.

https://www.kaggle.com/datasets/veeralakrishna/relational-strategies-in-customer-servicersics

Methodology:

The CRISP-DM framework will guide the project through its life cycle:

- **Business Understanding:** Define objectives through stakeholder consultation and develop domain knowledge.
- **Data Understanding:** Acquire and explore relevant data to uncover initial insights.
- **Data Preparation:** Clean and transform the data into a suitable format for modeling.
- **Modeling:** Select and train appropriate models aligned with the problem statement.
- **Evaluation:** Assess model performance using standard metrics; retrain with adjusted datasets if necessary.
- **Deployment:** Deploy the final model into a production environment where it can assist representatives during customer interactions.

Techniques and Tools:

The model will be developed using Recurrent Neural Networks (RNNs) implemented with TensorFlow and Keras.

Research Questions & Ethical Considerations

Research Questions:

1. Can a model accurately predict the next word in a customer service conversation given the historical context?
2. How can predictive text improve response time and service quality in customer support environments?

3. What types of responses are most commonly predicted, and how do they vary across different domains (e.g., travel vs. telecom)?

Ethical Implications:

- Data Privacy:

  All personally identifiable information (PII) in the dataset has been masked to protect customer identities. Numeric PII is replaced with '#' and text-based identifiers with placeholders like 'cname' and 'pname'.

- Bias Mitigation:
  - The data has been reviewed to minimize bias related to gender, location, or language.
- Responsible AI Use:
  - Care will be taken to ensure that the model does not reinforce harmful stereotypes or produce inappropriate suggestions. The tool is meant to support—not replace—human decision-making.

References

Next IT. (2016). *Relational Strategies in Customer Service (RSiCS) Dataset*. https://nextit-public.s3-us-west-2.amazonaws.com/rsics.html

TensorFlow. (n.d.). *Recurrent neural networks (RNN) with Keras*. TensorFlow. https://www.tensorflow.org/guide/keras/rnn

Didado, J. (n.d.). *Ethics in natural language processing*. DiDaDo. https://dida.do/blog/ethics-in-natural-language-processing