**Next Word Prediction-Language Model**

* **Topic**:

Language modeling involves predicting the next word in a sequence given the sequence of words already present. A language model is a key element in many natural language processing models such resolving customers inquiries through chat or answering the questions through emails.

* **Business Problem:**

In customer Service business especially in messaging or Chats or email supports, customer representative often struggle to response fast if they have limited knowledge of business area wherein the inquiry is about and need respond fast for better service and improved customer satisfaction.

Business Stakeholder wanted to build model that would learn from previously provided chat or email resolution history and suggest the next word when representative start providing resolution to the customers inquiry.

* Build the model that would predict the next word based on previous context.
* **Datasets**:

Relational Strategies in Customer Service (RSiCS) Dataset

Human-computer data from three live customer service Intelligent Virtual Agents (IVAs) in the domains of travel and telecommunications were collected, and annotators marked all text that was deemed unnecessary to the determination of user intention.

Data was collected from four sources. The conversation logs of three commercial customer service IVAs and the Airline forums on TripAdvisor.com during August 2016.

Dataset numbering used in files:

* TripAdvisor.com airline forum
* Train travel IVA
* Airline travel IVA
* Telecommunications support IVA
* **Methods**:

To build this model, CRISP-DM methodology will be used and followed each stage from it to make sure right product is built with minimal issues.

* Business Understanding - Business objective will be determined by discussing with stakeholders and respective knowledge about domain will be gained to learn more about dataset and its sources.
* Data Understanding – Relevant data will be collected, and exploratory data analysis will be performed to gain more understanding from the data.
* Data Prep – In this stage data will be cleaned and transformed into format that ML model would accept it.
* Modeling – Model would be finalized based on problem statement and model will be trained and validated with training data.
* Evaluation – Trained model will be evaluated based on different metrics and result will be reviewed and if needed dataset will be changed and model will be re-trained with new changed datasets.
* Deployment – The final model will be deployed into production system, where it would help representatives to predict next word when they chat with customer or respond resolution in the form of email.
* Techniques – The model would be built using Recurrent neural networks (RNN) and TensorFlow and Keras.
* **Ethical Considerations:**

All the personal identifiable information (PII) present in the datasets have been removed by tokenizing it. All numerical characters contained PII such as account number SSN, phone numbers are masked with ‘#’ and alphabetic characters such as name, address, company name are masked with ‘cname’ and ‘pname’.

This data is acquired in August 2016, made sure that it not biased based on gender or specific geographic area. Model would be trained purely from the historical data and will make sure there would not be any ethical bias.

* **References**:
* <https://nextit-public.s3-us-west-2.amazonaws.com/rsics.html>
* <https://www.tensorflow.org/guide/keras/rnn>
* <https://dida.do/blog/ethics-in-natural-language-processing>