

# INTENT CLASSIFICATION – PROJECT PROPOSAL

## COMPUTATION LINGUISTICS AND ANALYSIS; CSCI-B 590, SPRING 2022

ABHIJIT NAYAK (nayakab), ABHISEK PANIGRAHI (abpani)

**Introduction:** Intent Classification is the automatic categorization of text data based on customer goals. It is known to be a complex problem in Natural Language Processing (NLP) research. This problem represents one of the stepping stones to obtain machines that can understand our language. This technique has various applications one of which helps businesses grow. For example, identifying a potential customer's intent to purchase early on, then there will be a better chance of converting them to fully fledged customers. Intent classification uses machine learning and natural language processing to automatically associate words or expressions with a particular intent. For example, a machine learning model can learn that words such as buy or acquire are often associated with the intent to Purchase. If you're analyzing customer emails, you might choose the tags *Interested*, *Need Information*, *Unsubscribe*, *Wrong Person*, *Email Bounce*, *Autoreply*, etc.

### **Collected Datasets:**

- 1) **Out of Scope Intent Classification Dataset:** "Out-of-scope" inputs are those that do not belong to the set of "in-scope" target labels. There are 150 in-scope intents (training set), and the intent classification is to be done on out-of-scope data (testing set).
- 2) **Chatbot Intent Classification Dataset:** This is a dataset containing chatbot conversations and intent classification is to be performed on this. The results can be helpful when creating chatbots or conversational interfaces.
- 3) **Restaurant Intent Entities:** This is a conversation dataset, mainly containing conversations between customers and the restaurant executives. Intent classification would be performed on this domain specific dataset.

**Research Question 1:** Performing linguistic analysis on the corpus like keyword analysis, lexical trends.

**Research Question 2:** Comparing different word embeddings: from static to contextual

**Research Question 3:** Performing a downstream task sequence labelling after the intent classification (possibly to make the named entity recognition process easier in question/answering or conversation type datasets)

We will be using SVM, Random Forest or pre-trained models to train the NLU models as well as Word2Vec and Sentence Bert in static and contextual embeddings respectively. There would be some challenges during this project like classifying intents on out-of-scope data and/or classifying multiple intents. Challenges would be there in performing the linguistic analysis as accurately as possible like what exact methods to follow.