Final Presentation – Classification of Text ML-Approach

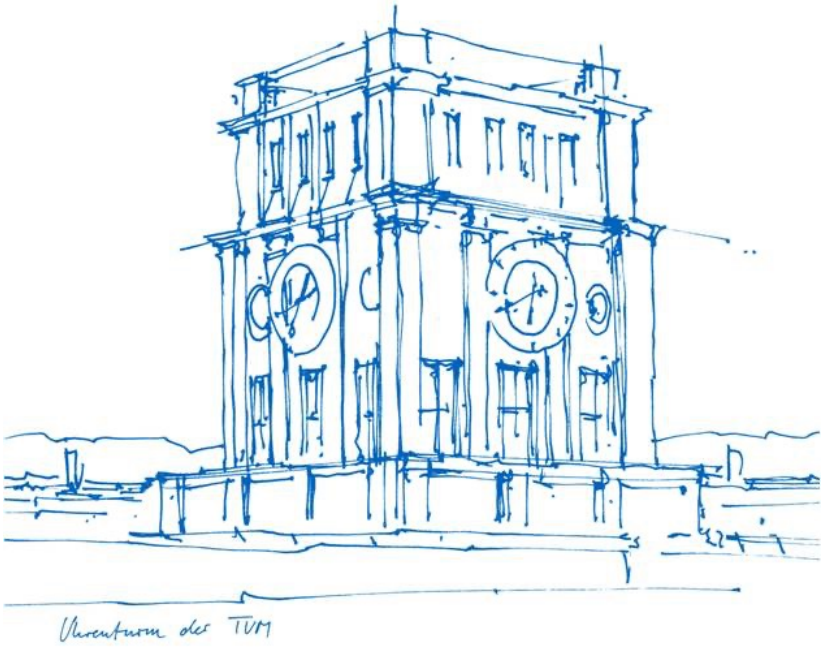
Master Internship: Approaching Information System Challenges with Natural Language Processing (IN2106, IN2130)

Patrick Ahrend

Garching, 30 January 2024

Agenda

1) Mission
2) How I planned the project
3) How it went
4) Explanatory UMAP
5) Results
6) Code Demo



Mission: Situation – Complication – Resolution

Situation

Companies have lots of business processes which need to be affirmative to legal requirements

Complication

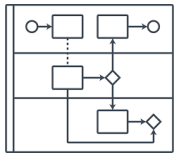
The process of legal work is quite labour-intensive and costly.

Solution

Semi Automation with ML model which is able to identify new regulations for human review

Solution: Machine Learning Pipeline for text classification

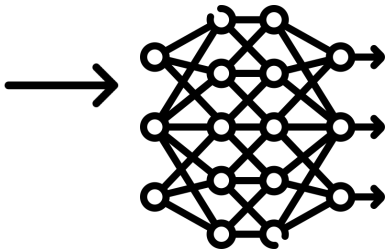
Input



Process Description

Legal text

ML Pipeline



Output

{1}

{1}

{0}

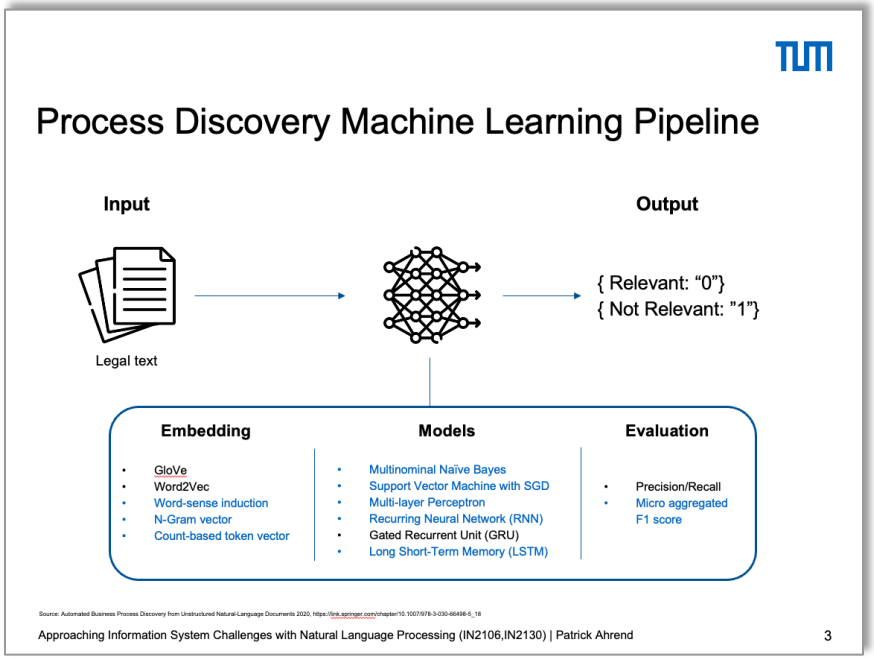
0= Irrelevant
1= Relevant
For the process

Data used

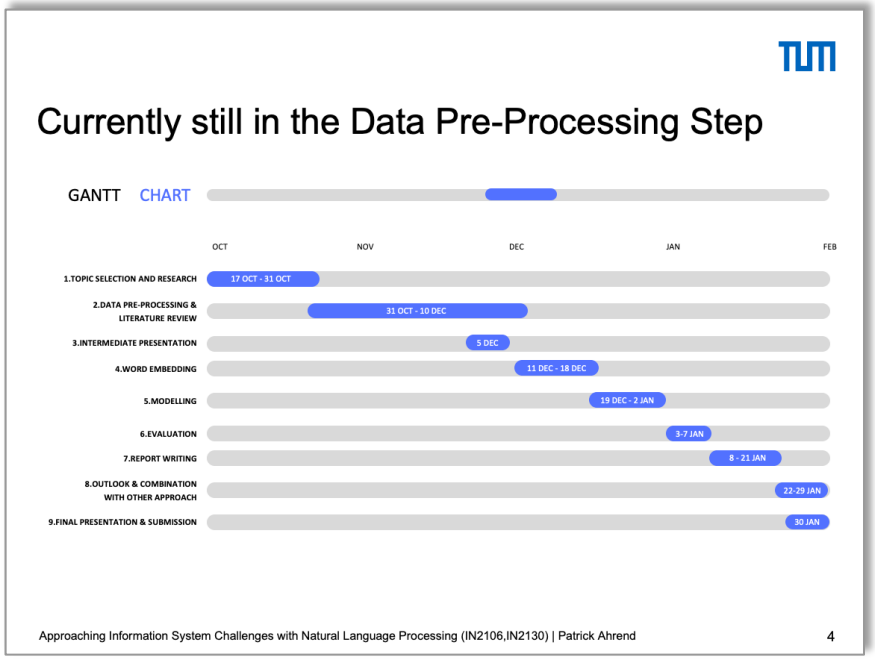
- 3 data sources: Smart Meter, Australia Data and GDPR
- 17 process in total
- 1860 data points
- {Process Description, Legal Text, Label}

How the project was planned

Project Goal

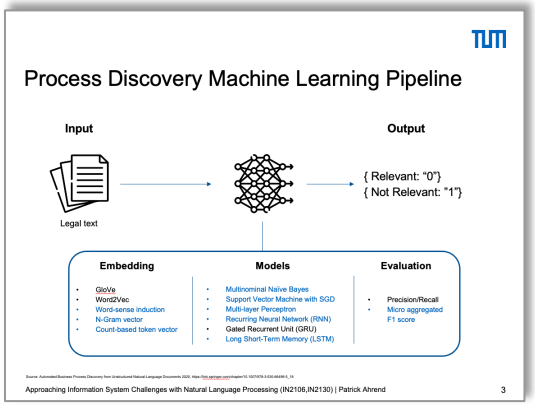


Project Timeline



How the project actually happenend

Project Goal



GloVe

Word2Vec

fastText

Bert

TF-IDF

Decision Tree

Logistic Regression

Gradient Boosting

Naïve Bayes

Perceptron

Random Forest

SVC

SDG

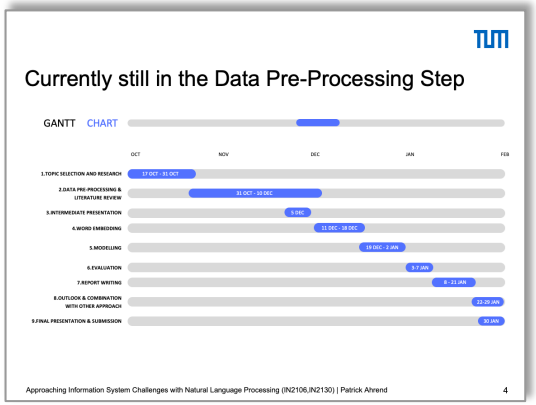
scikit learn

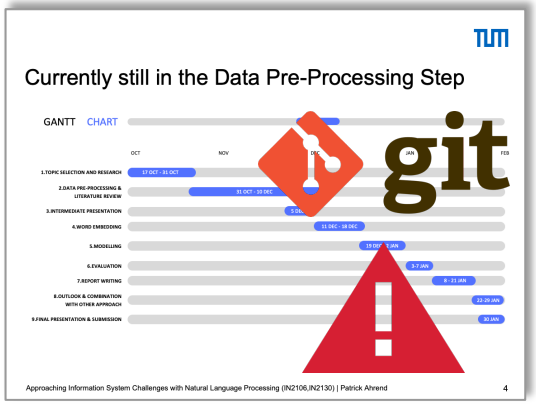
Bert

RNN

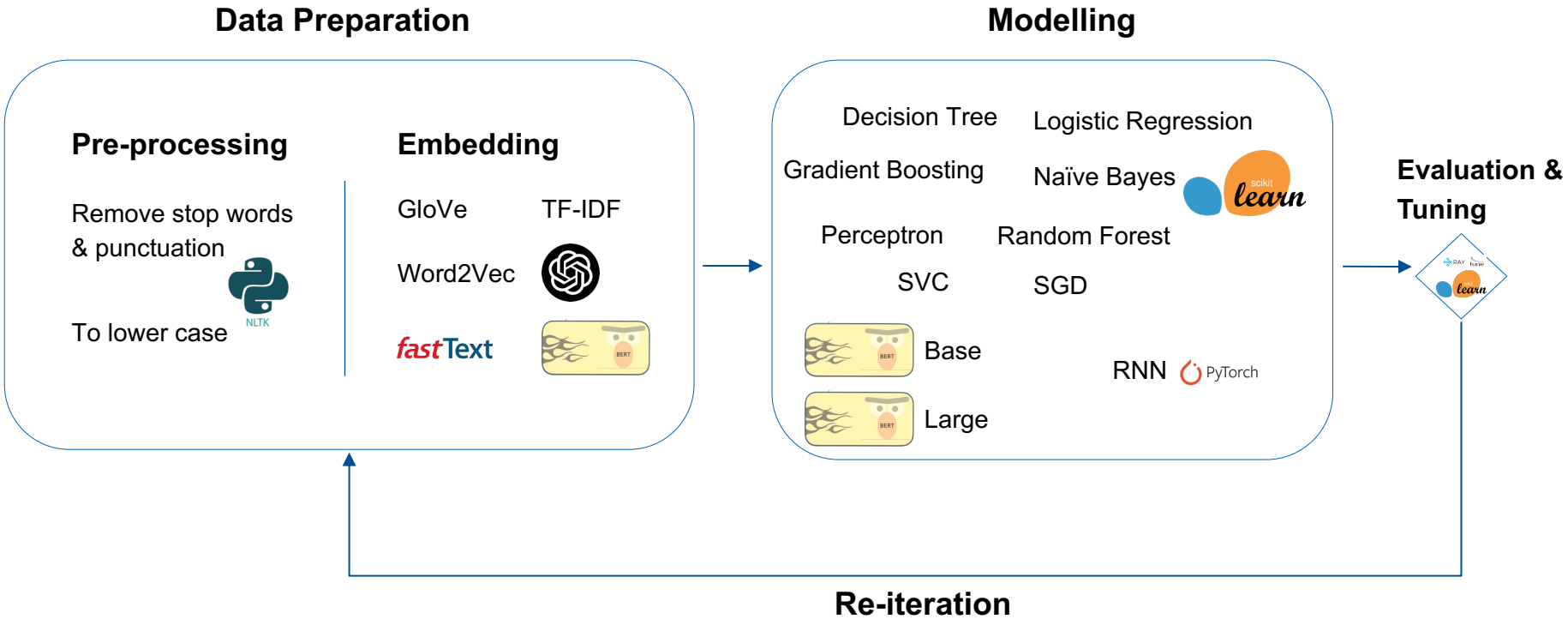
PyTorch

Project Timeline

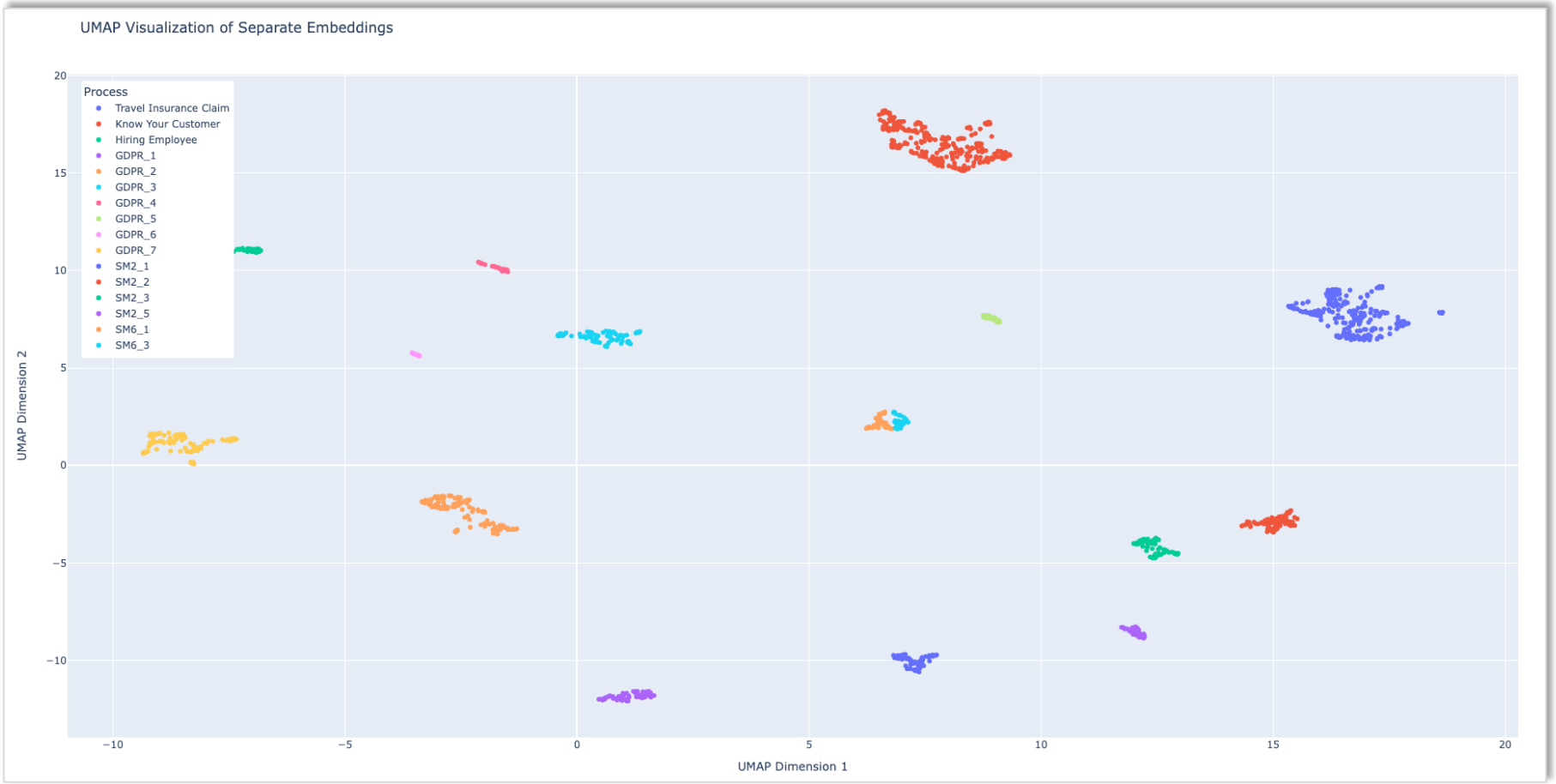




Machine Learning Pipeline for text classification

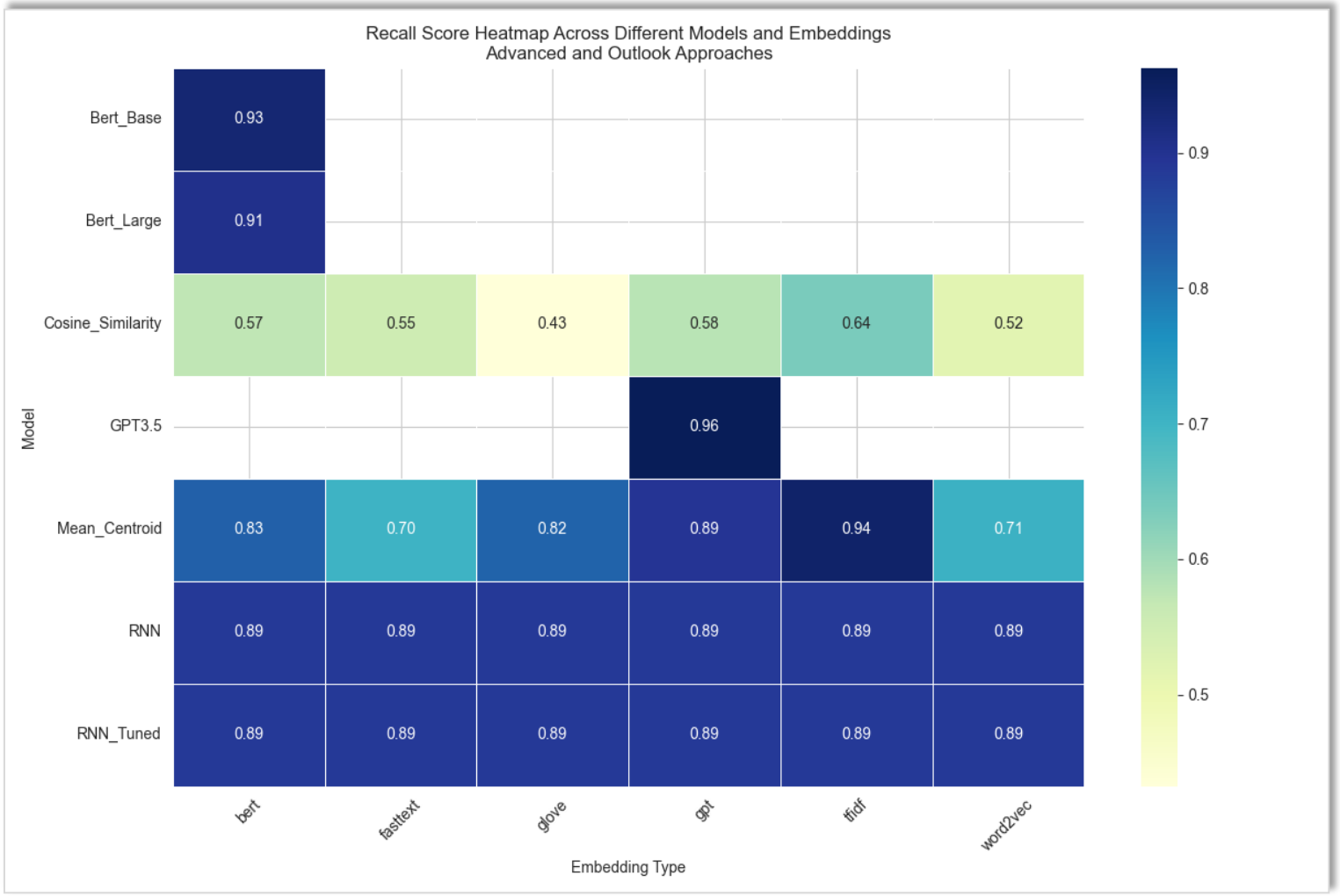


UMAP for Explanatory Data Analysis



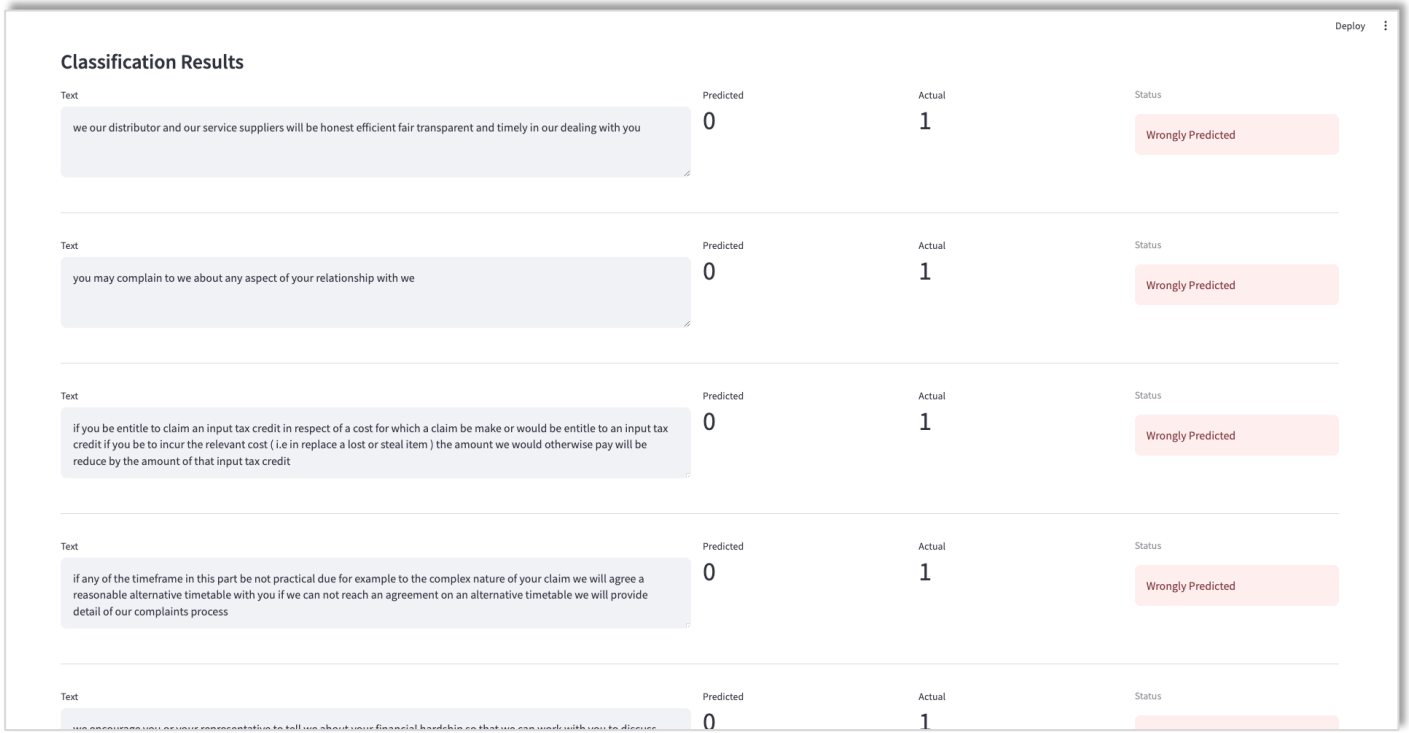
Evaluation - Quantitatively I/II

Evaluation - Quantitatively II/II



Evaluation - Qualitatively

- **9** data points from travel insurance
- SGD, RNN, Perceptron, SVD failed to learn
- GPT and Bert predicted quite differently



Code Demonstration

Thanks!
Questions/Feedback?

Implementation

UML of the pipeline
?

- Qualitative
- Quantitative