Capstone Project

TICKET CLASSIFICATION

Team: Mokshada Joshi, Mukesh Joshi & Subhash Kumar

Hero Vired | A HERO GROUP COMPANY

BUSINESS MOTIVATION

- Automating ticket assignment process to reduce response time for customer queries.
- Targeting a reduction in first effective response time by 18 hours on average.
- Utilizing data-driven approaches to classify customer tickets accurately.
- Leveraging the dataset from the customer support desk for analysis and modelling.
- Improving overall customer satisfaction by resolving queries more efficiently.
- Enhancing operational efficiency and resource utilization within the bank.
- Streamlining communication channels between customers and relevant stakeholders.
- Potentially reducing workload and stress on support staff by automating ticket assignments.
- Aim for cost savings through improved efficiency and reduced manual intervention.

APPROACH

The entire work has been approached in three tasks:

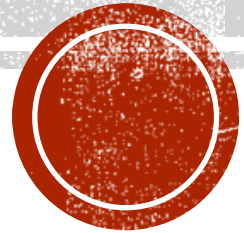
Task 1: Data Understanding

Task 2: Feature Engineering

Task 3: Modelling

Task 4: Deployment

TASK 1: DATA UNDERSTANDING

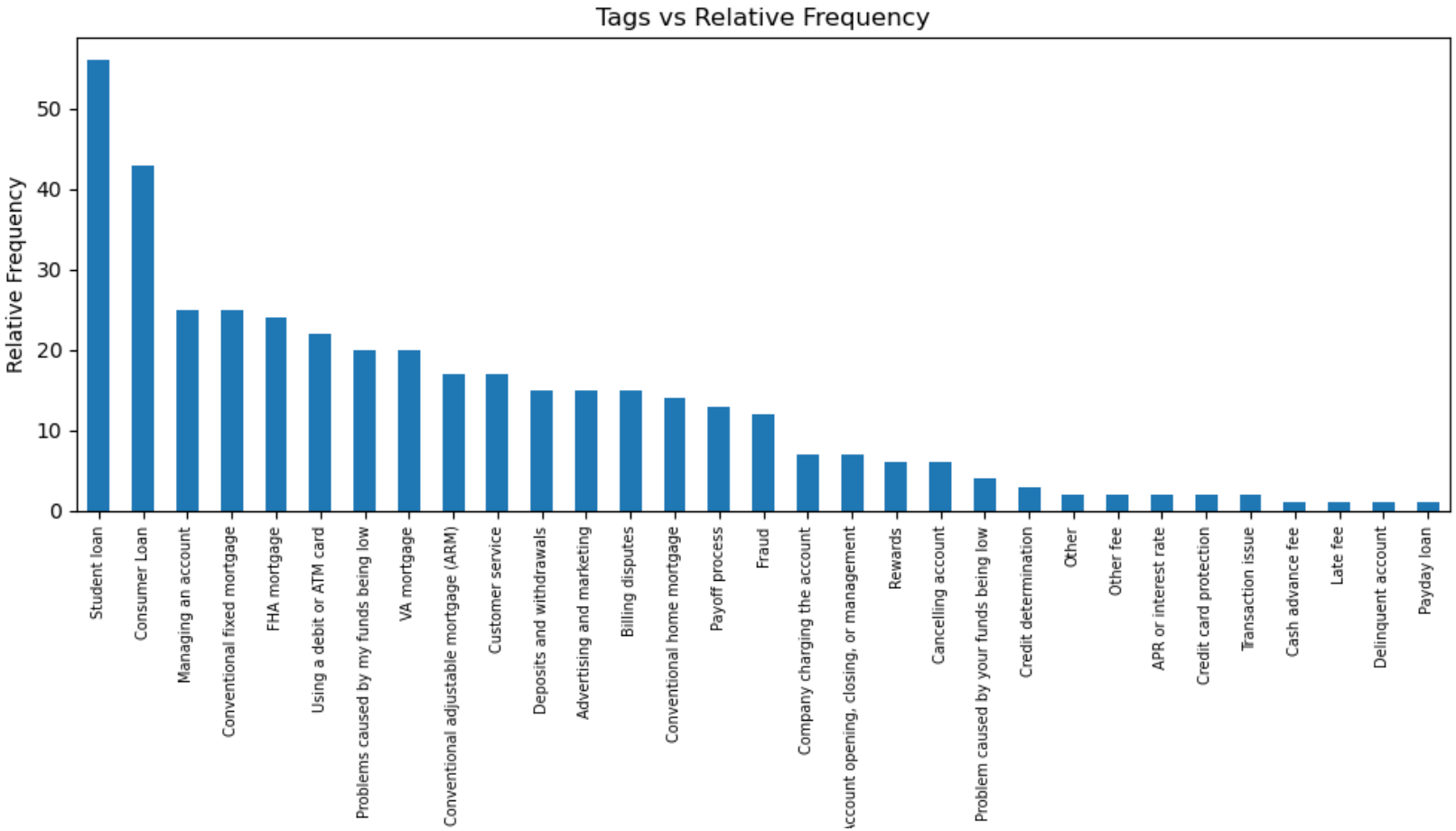


DATA UNDERSTANDING

- Number of labelled customer complaints: 400

DATA UNDERSTANDING

- Relative frequency of tags in the dataset



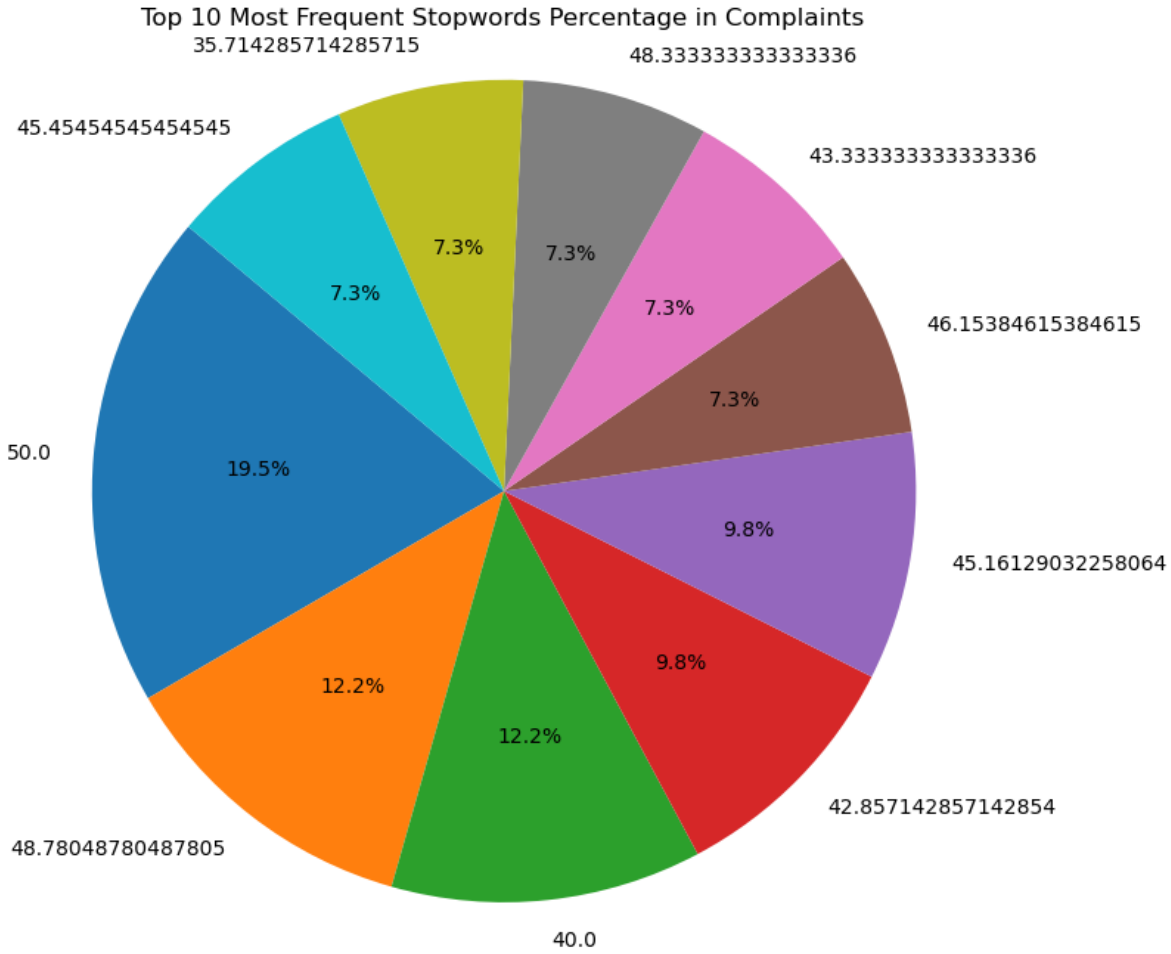
Tags	Frequency
Student loan	0.1400
Consumer Loan	0.1075
Managing an account	0.0625
Conventional fixed mortgage	0.0625
FHA mortgage	0.0600
Using a debit or ATM card	0.0550
Problems caused by my funds being low	0.0500
VA mortgage	0.0500
Conventional adjustable mortgage (ARM)	0.0425
Customer service	0.0425
Deposits and withdrawals	0.0375
Advertising and marketing	0.0375
Billing disputes	0.0375
Conventional home mortgage	0.0350
Payoff process	0.0325
Fraud	0.0300
Company charging the account	0.0175
Account opening, closing, or management	0.0175
Rewards	0.0150
Cancelling account	0.0150
Problem caused by your funds being low	0.0100
Credit determination	0.0075
Other	0.0050
Other fee	0.0050
APR or interest rate	0.0050
Credit card protection	0.0050
Transaction issue	0.0050
Cash advance fee	0.0025
Late fee	0.0025
Delinquent account	0.0025
Payday loan	0.0025

DATA UNDERSTANDING

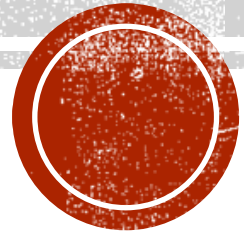
- Checking for encoding issues:
  - File read successfully with encoding: utf-8
  - File read successfully with encoding: latin1
  - Error reading file with encoding utf-16: 'utf-16-le' codec can't decode byte 0x0a in position 287344: truncated data
  - File read successfully with encoding: iso-8859-1

DATA UNDERSTANDING

- Stopwords in each complain:
  - Minimum stopwords %: 29.0%
  - Maximum stopwords %: 50.0%
  - Mean stopwords %: 37.815%



TASK 2: FEATURE ENGINEERING



FEATURE ENGINEERING

- Count matrix or a tf-idf matrix
  - Shape of TF-IDF matrix: (400, 3772)

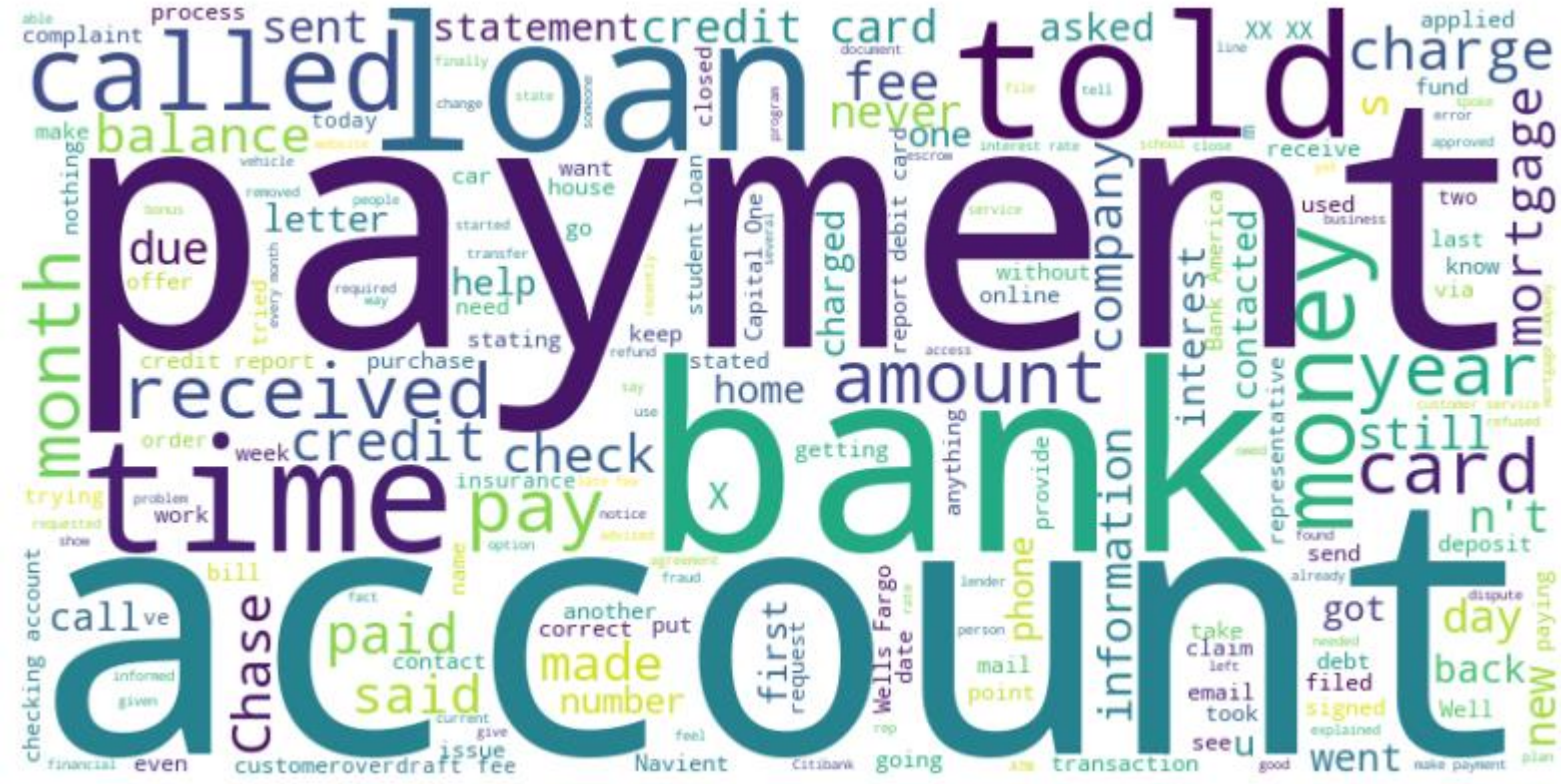
FEATURE ENGINEERING

- Count matrix or a tf-idf matrix
  - Adding other features apart from TF-IDF matrix: 1. Sentiment Score and 2. Sentiment Level

FEATURE ENGINEERING

- Sentiment analysis: Creating Word Clouds to visualize the most frequent words in complaints

As per Unigram analysis, data revolve around Payment, Account, Loan



FEATURE ENGINEERING

- Sentiment analysis: Creating Word Clouds to visualize the most frequent words in complaints

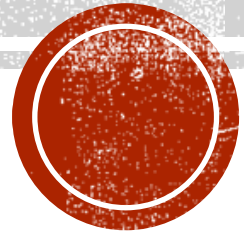
As per Bigram analysis, data revolve around Credit Card, Credit report, Checking account

FEATURE ENGINEERING

- Sentiment analysis: Creating Word Clouds to visualize the most frequent words in complaints

As per Trigram analysis, data revolve around xx xx 2016, 00 per month, jp morgan chase

TASK 3: MODELING



MODEL BUILDING

- Models explored to create a ticket classification system are:
  - Naive Bayes
  - Linear Classifier
  - SVM
  - Tree-based ensembles
  - KNN
  - Bert Sequence Classifier

MODEL BUILDING

- Model Performance Evaluation

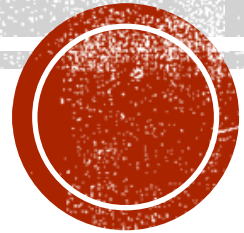
Models	Recall	Precision	Accuracy	Silhouette score -
Naive Bayes	22	25	35	0.00
KNN	21	18	33	0.00
SVM	24	25	40	0.00
LSTM	0	0	22	0.00
K Means	0	0	0	0.80
Bert Sequence	29.3	20	43	0.00

As per the model performance matrix, the Bert Sequence Classifier shows the highest accuracy.

MODEL BUILDING

- Remarks on Model Performance Evaluation:
  - The aim of the project is to provide an initial insight into the category of the complaint received.
  - Here, the size of the data set is only 400 rows. So, to have better results more data points are required.

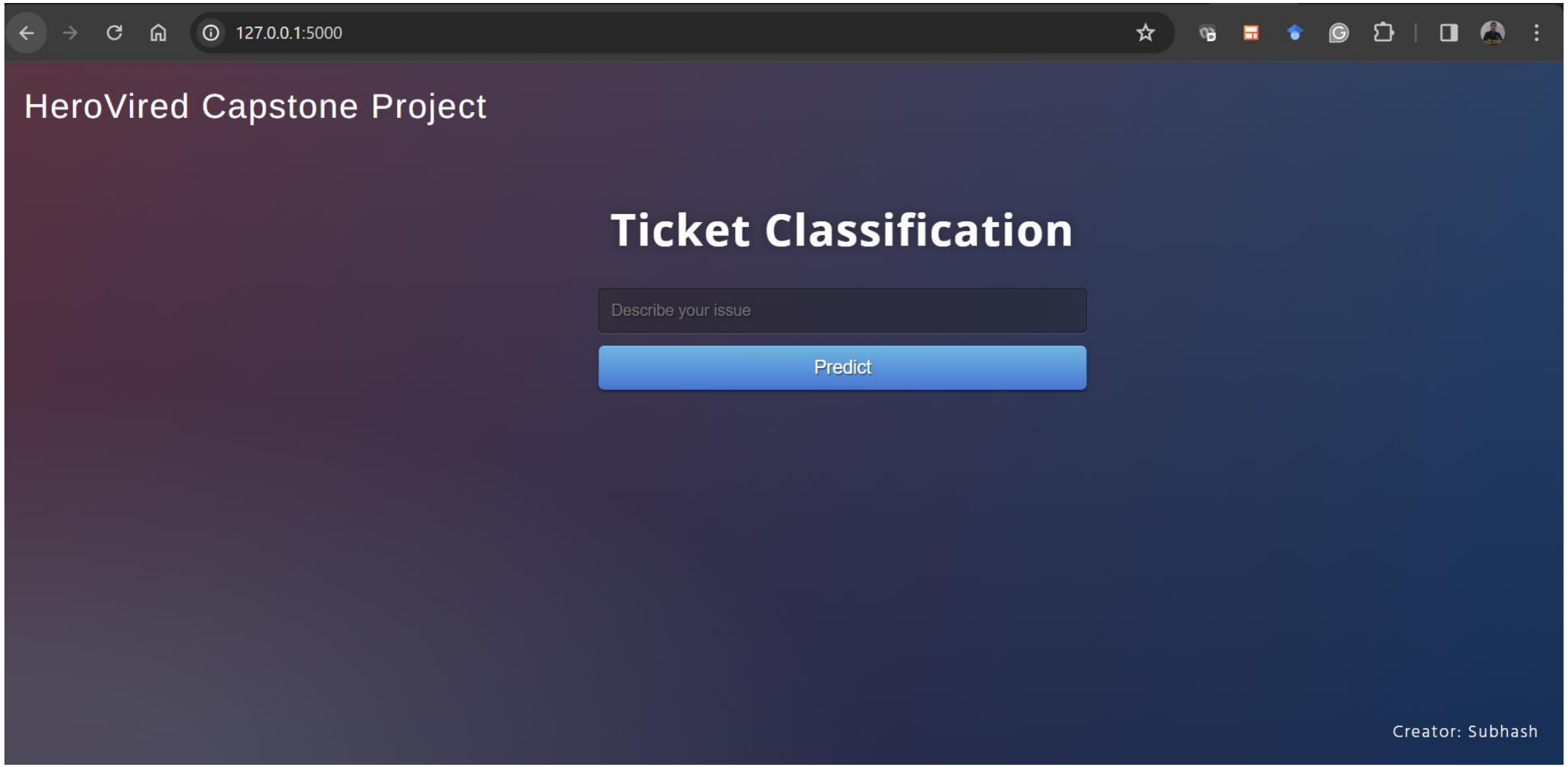
TASK 4: MODEL DEPLOYMENT



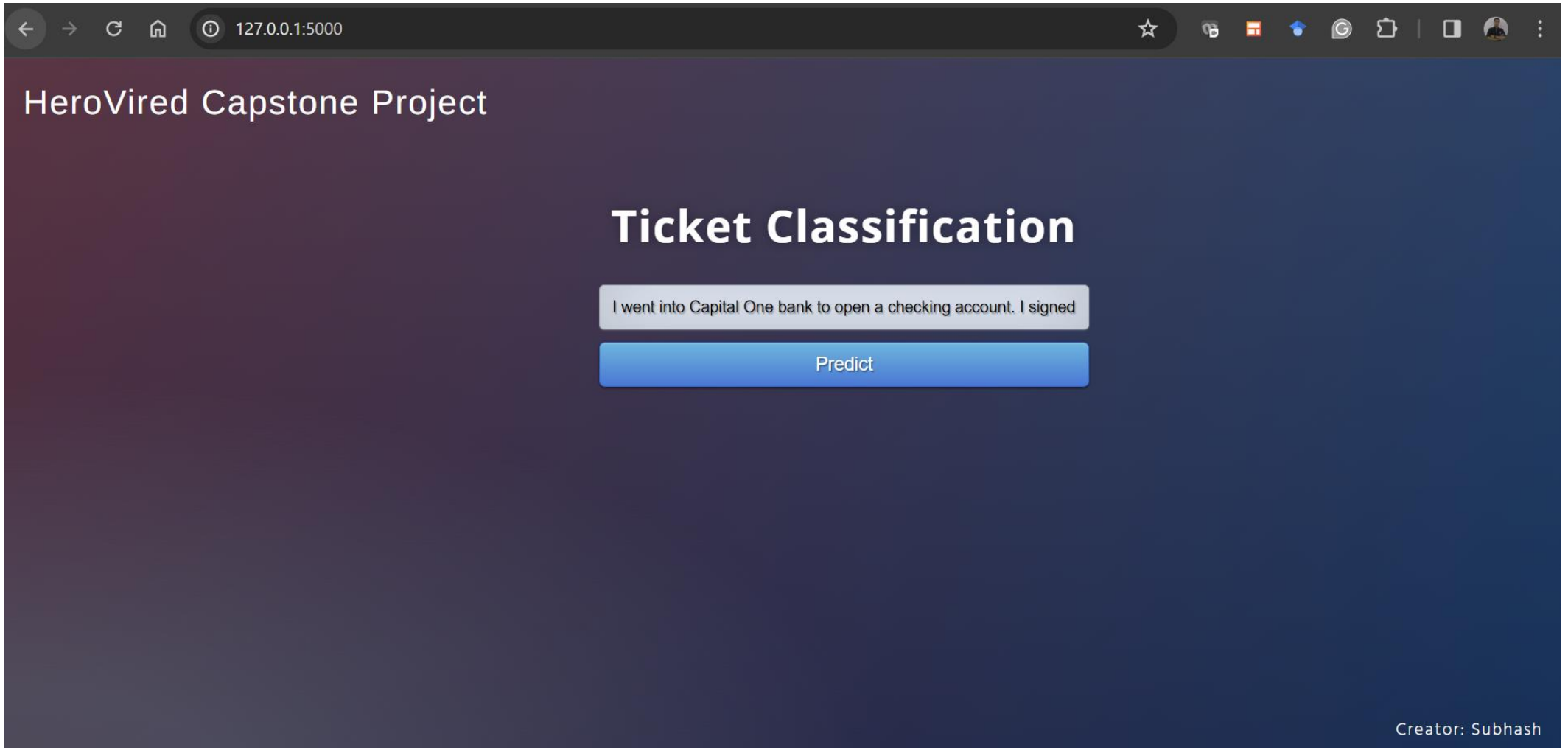
PREDICTION MODULE DIRECTORY

static	14/03/2024 15:50	File folder	
templates	14/03/2024 15:50	File folder	
app	16/03/2024 19:22	PY File	2 KB
complaints	16/03/2024 16:53	Microsoft Excel Comma Separated Values File	273 KB
fine_tuned_bert_model.pkl	16/03/2024 01:06	PKL File	427,864 KB
label_encoder.pkl	16/03/2024 01:06	PKL File	1 KB
model	16/03/2024 16:57	PY File	6 KB
README.md	14/03/2024 15:50	MD File	1 KB
Ticket Classification-final	16/03/2024 01:11	Microsoft PowerPoint Presentation	1,011 KB
ticketClassifier	14/03/2024 16:05	Microsoft Excel Comma Separated Values File	124 KB

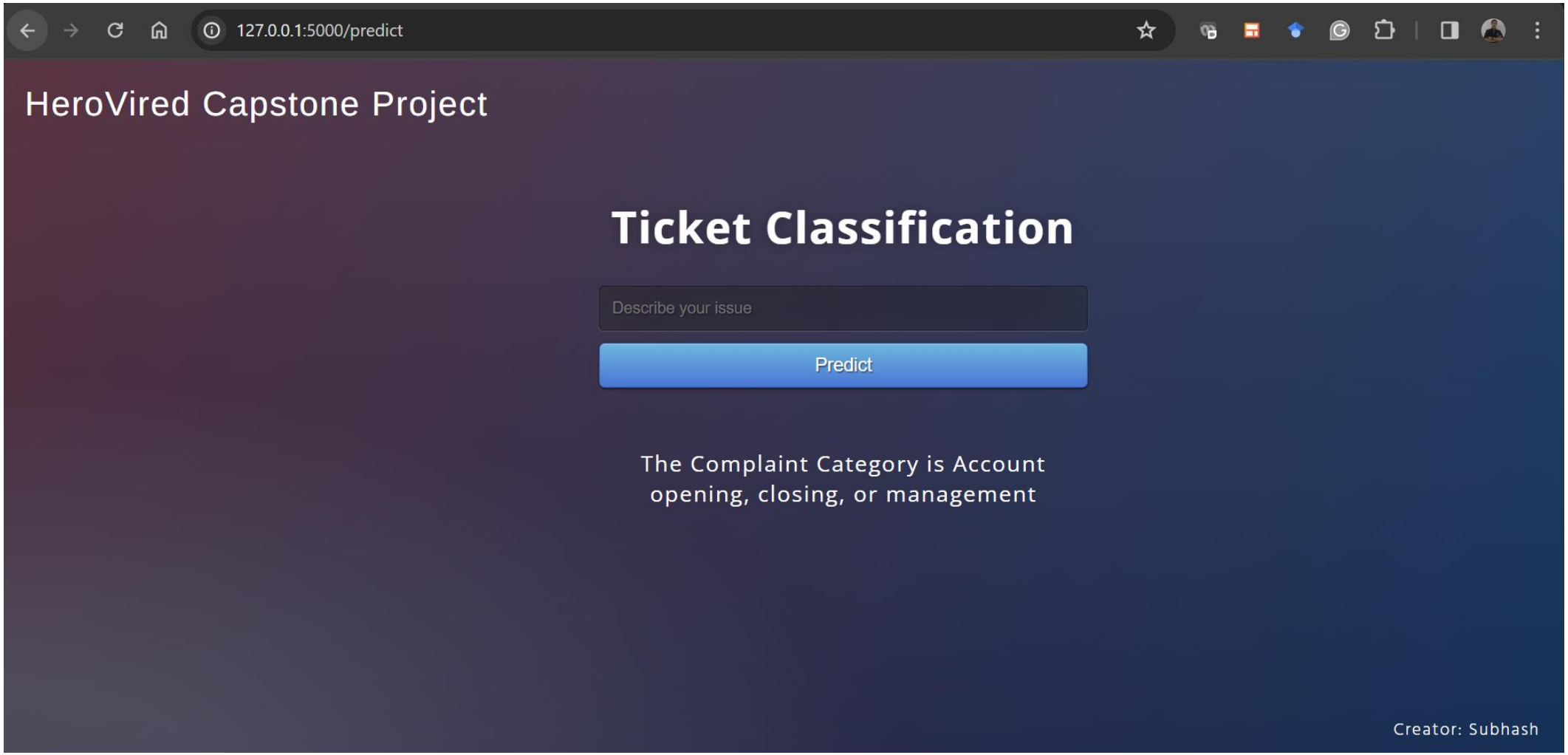
LOCAL HOST FOR CHECKING SPECIFIC COMPLAINTS



ENTERING COMPLAINTS IN TICKET CLASSIFICATION CHECK BOX



PREDICTING TYPE OF COMPLAINT



THANK YOU!