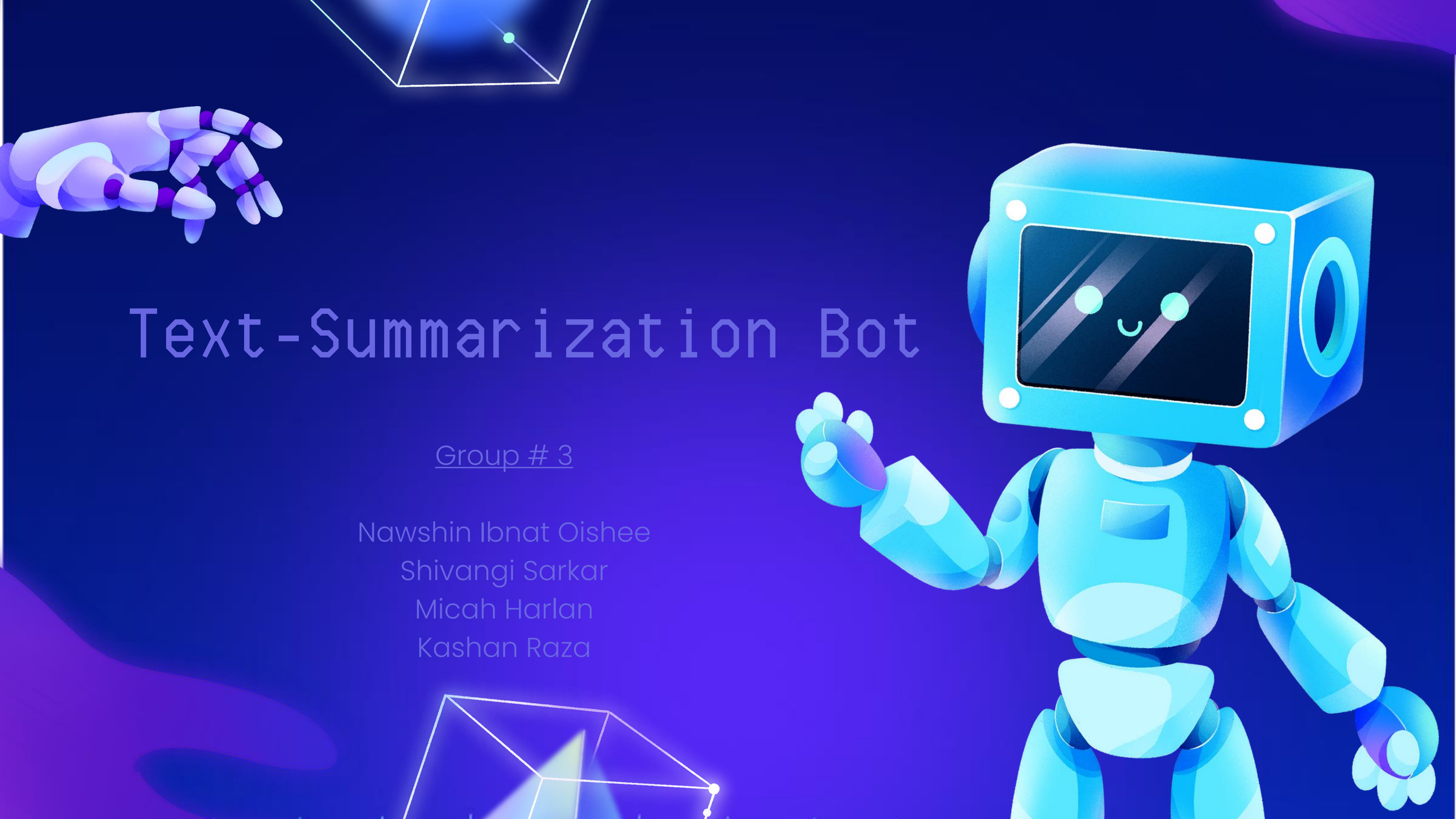


Text-Summarization Bot

Group # 3

Nawshin Ibnat Oishee
Shivangi Sarkar
Micah Harlan
Kashan Raza

TABLE OF CONTENTS

TEAM INTRODUCTION

Nawshin Ibnat Oishee
- MEng in CS, Virginia Tech
- Concentration in Data Analytics & AI

Shivangi Sarkar
- MEng in CS, Virginia Tech
- Concentration in Data Analytics & AI

Micah Harlan
- MEng in CS, Virginia Tech
- Concentration in Data Analytics & AI

Kashan Raza
- MEng in CS, Virginia Tech
- Concentration in Cybersecurity

Problem Statement

Problem

we aim to address is the overwhelming volume of information, specifically textual data in the form of articles, research papers, and documents, that are available to us which can make it challenging for readers to quickly extract key insights and relevant information efficiently

Significance

This project aims to address the need to simplify and accelerate the process of extracting valuable information from complex pieces of literature by using AI-powered text summarization techniques, making research more accessible and time-efficient. Solving this problem would decrease the amount of time it takes researchers to review and critically analyze research papers within their field of study, aiding in peer reviews, and the general study of complex research. Overall, this would make academic journals (Along with any other category of literature the user wishes to summarize) more accessible for students, researchers, and the general public.

Use-Case Scenarios

01

A student may need to complete an assignment, involving examining various research papers. Reading large bodies of text can be time-consuming and stressful for a student who needs to meet a deadline. The use of a text summarization tool would be helpful and efficient, allowing the student to find the right paper in a shorter time. They would then be able to complete their assignment sooner.

02

A professional data analyst working at a large company is tasked with creating a new tool based on recent research in a related field. The professional is not a researcher and isn't familiar with the particular research, they find the academic papers difficult to read. Research papers are known to be lengthy and complex, often because of the scrutiny they are put under during peer reviews, but this makes them unattainable for the general public to understand without prior knowledge. This professional could use our AI tool to summarize the relevant research and understand it just enough to create his tool. They don't need to understand every mathematical formula or critique the paper, so this level of understanding is sufficient for them.

AI Algorithm and Model

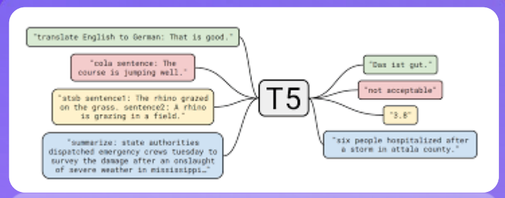
Discord Bot

- A popular social platform with programming integration through its Bots API

Hugging Face

- Hugging Face is a machine learning (ML) and data science platform known for its open-source AI models and natural language processing (NLP) tools.

Fine-Tuned T5-Base Language Model



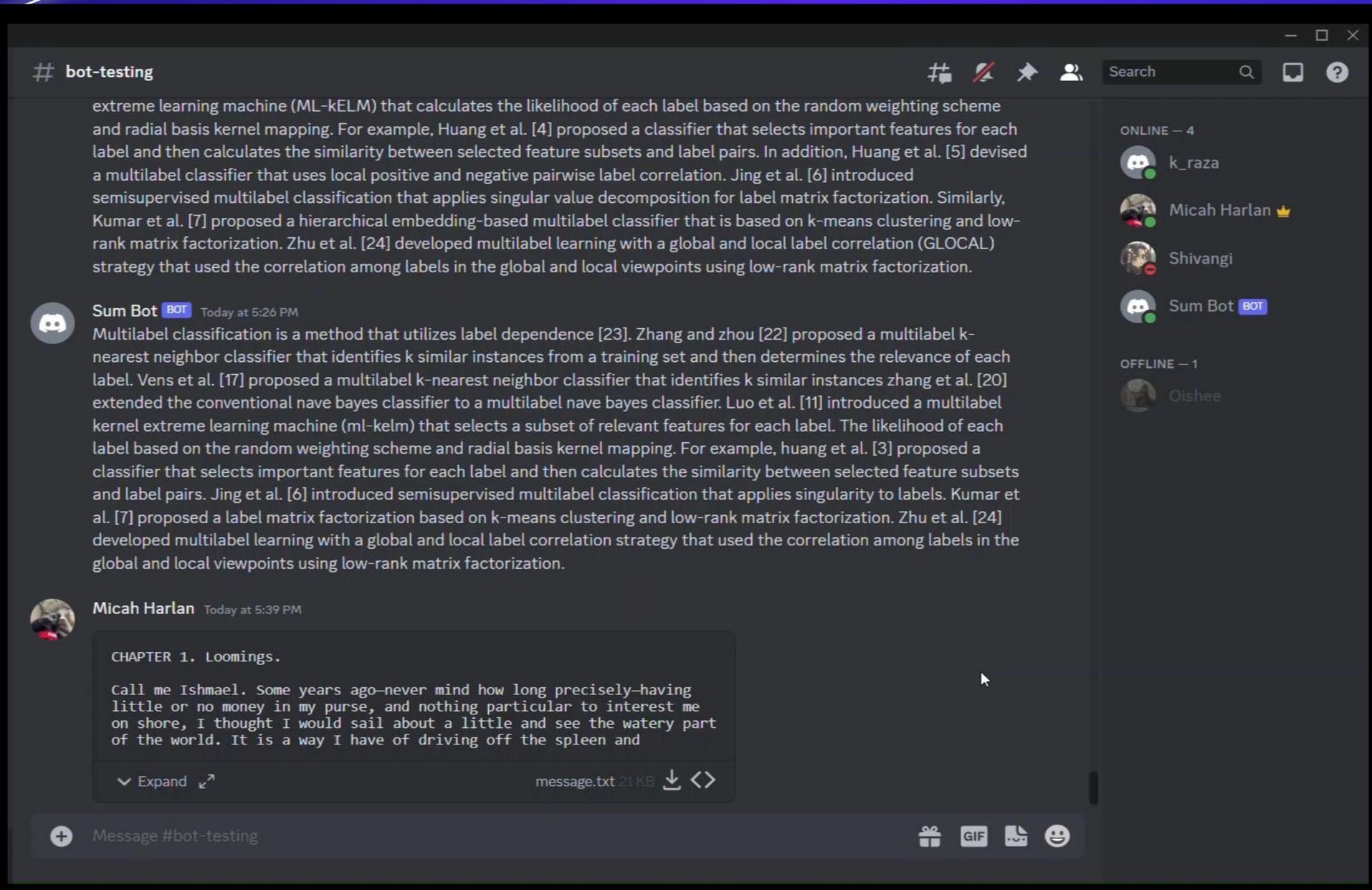
- An encoder-decoder model trained on supervised and unsupervised tasks designed for text-to-text conversion
- Supervised training uses downstream tasks provided by the GLUE benchmark
- Works well for summarization, translation, and other NLP tasks

Results and Demonstration

01

02

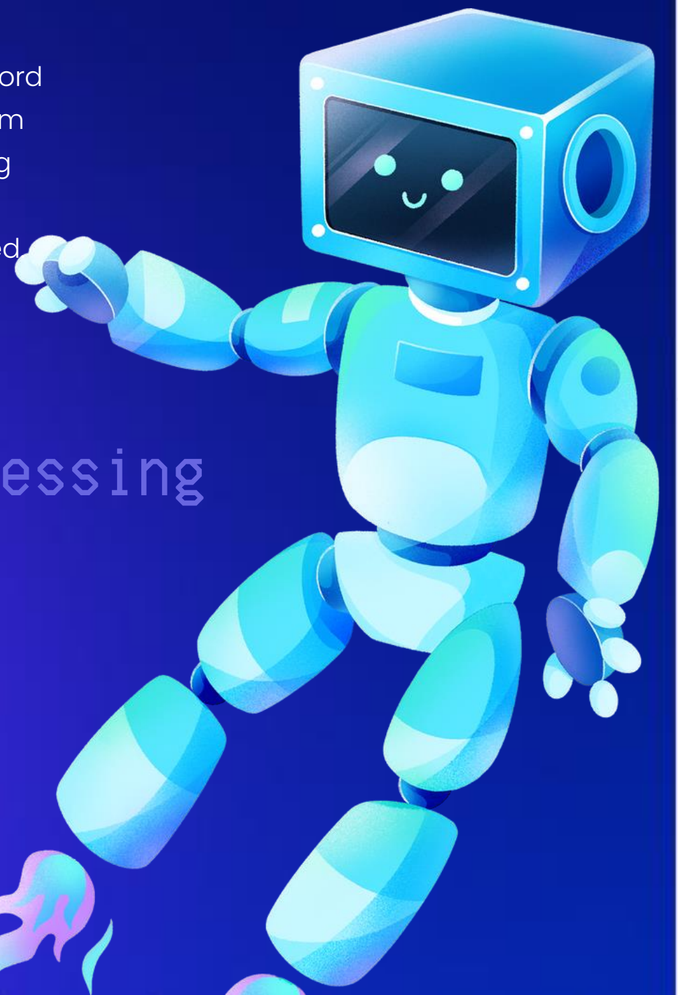
- We successfully created our summarizer Discord bot
- Limitations:
  - Due to Discord's 2000-character limit for non-subscribed users, we decided to create a Python script to accompany our bot
    - This script handles cases of larger text
    - Ideal for summarization of large papers
  - Summaries are not perfect
    - Model is pre-trained
    - Would need fine-tuning with datasets containing summaries of research papers


bot-testing
Search
extreme learning machine (ML-kELM) that calculates the likelihood of each label based on the random weighting scheme and radial basis kernel mapping. For example, Huang et al. [4] proposed a classifier that selects important features for each label and then calculates the similarity between selected feature subsets and label pairs. In addition, Huang et al. [5] devised a multilabel classifier that uses local positive and negative pairwise label correlation. Jing et al. [6] introduced semisupervised multilabel classification that applies singular value decomposition for label matrix factorization. Similarly, Kumar et al. [7] proposed a hierarchical embedding-based multilabel classifier that is based on k-means clustering and low-rank matrix factorization. Zhu et al. [24] developed multilabel learning with a global and local label correlation (GLOCAL) strategy that used the correlation among labels in the global and local viewpoints using low-rank matrix factorization.
Sum Bot BOT Today at 5:26 PM
Multilabel classification is a method that utilizes label dependence [23]. Zhang and zhou [22] proposed a multilabel k-nearest neighbor classifier that identifies k similar instances from a training set and then determines the relevance of each label. Vens et al. [17] proposed a multilabel k-nearest neighbor classifier that identifies k similar instances zhang et al. [20] extended the conventional nave bayes classifier to a multilabel nave bayes classifier. Luo et al. [11] introduced a multilabel kernel extreme learning machine (ml-kelm) that selects a subset of relevant features for each label. The likelihood of each label based on the random weighting scheme and radial basis kernel mapping. For example, huang et al. [3] proposed a classifier that selects important features for each label and then calculates the similarity between selected feature subsets and label pairs. Jing et al. [6] introduced semisupervised multilabel classification that applies singularity to labels. Kumar et al. [7] proposed a label matrix factorization based on k-means clustering and low-rank matrix factorization. Zhu et al. [24] developed multilabel learning with a global and local label correlation strategy that used the correlation among labels in the global and local viewpoints using low-rank matrix factorization.
Micah Harlan Today at 5:39 PM
CHAPTER 1. Loomings.
Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and
Expand
message.txt 21 KB
Message #bot-testing
GIF
ONLINE — 4
k_raza
Micah Harlan
Shivangi
Sum Bot BOT
OFFLINE — 1
Oishee

CONTINUOUS LEARNING

Discord Bot Creation

Most of us were unfamiliar with creating our own Discord bots and the requirements for connecting our program to Discord. We quickly found that although connecting to Discord was easy, formatting our bot's output was challenging and required more time than we expected.



Natural Language Processing

The creation of our bot required familiarity with existing NLP models and databases used for training. We spent a large portion of our design phase deciding on which model would be best suited for our project and integration with Discord. After plenty of research on existing NLP models, we decided to use the base version of T5, which has already been tested and used for text summarization.

THANK YOU!