## **Assignment - 5**

## 

**Deploying Hate Speech Detector**

Northeastern University: College of Professional Studies

EAI6010: Applications of Artificial Intelligence

**Professor**: Sergiy Shevchenko

Anjal Mohammed

June 24th, 2025

## 

## **Introduction**

This microservice is a web application that enables users to identify hate or toxic speech in short texts like tweets or social comments. The microservice utilizes Longformer, a pre-trained BERT model from Hugging Face: unitary/toxic-bert, fine-tuned to identify toxic content.

This service is implemented using Streamlit, a lightweight web framework for Python for quickly deploying ML apps. The model is loaded once on start-up and then used to classify the user-supplied text, as needed.

**General Input and Output**

**Input:**

* A short piece of text entered by the user (e.g., a tweet, message, or comment).
* The input must be a non-empty string, written in English.

**Output:**

* Label: Classification of the text, either **“toxic”** or **“non-toxic”**.
* Confidence Score: A probability score between 0 and 1 indicating the model’s confidence in its prediction.

**Input:** I hate you and your kind!

**Output:**

**Label:** toxic

**Confidence:** 0.96

**Input:** Hope you have a great day ahead! Stay positive.

**Output:**

**Label:** non-toxic

**Confidence:** 0.99

**How to Access and Use the Service**

**Live URL: https://hatespeechdetector-eai6010.streamlit.app/**

**Instructions:**

1. Go to the provided URL in any browser.
2. Enter your text in the input box labeled “Enter a tweet or text here”.
3. Click the “Classify” button.
4. Wait for the model to analyze the input and return the label and confidence score.

### **Conclusion**

This deployment demonstrates how an existing NLP model can be easily wrapped as a user-friendly web microservice. The interface is intuitive, and the backend model is robust enough for basic hate speech detection tasks. Future enhancements could include support for batch input, visualization of results, or multilingual detection.

**References:**

**Hsu, W.-H.** (n.d.). *Hate speech Twitter (Testing Set)* [CSV file]. Hugging Face. . Retrieved June 22, 2025, from <https://huggingface.co/datasets/thefrankhsu/hate_speech_twitter/resolve/main/testing%20set.csv>

**Streamlit. (n.d.).** *Streamlit documentation*. Streamlit, Inc. Retrieved June 22, 2025, from <https://docs.streamlit.io/>