HACETTEPE ÜNİVERSİTESİ

PROFANITY CHECK

Real-Time, On Device Turkish Speech Moderation

İrem Saygı **Sara Rzayeva**

Assoc. Prof. Ali Seydi Keçeli

ABSTRACT

We present a fully local, real-time system that listens to a microphone feed, transcribes Turkish speech with Whisper, pinpoints offensive words using a fine-tuned DistilBERT classifier, and instantly plays back a “clean” audio stream where detected profanities are replaced by adaptive beeps while the transcript is star-censored.

What is Profanity Check?

Real-time, on-device profanity filter.

- Listens to your mic, transcribes with Whisper.
- Flags Turkish swear words via a fine-tuned DistilBERT.
- Beeps out bad words and masks them in text–instantly.
- Runs 100 % locally for sub-second latency and full privacy.

Perfect for classrooms, meetings, streams, or any live talk.

Motivation

- Live streams, classrooms and public spaces in Türkiye need instant, privacy-preserving speech moderation.
- Cloud services add latency and leak data; we run fully local on a laptop-class GPU / CPU.
- We even censor words that might be misinterpreted as slurs, minimising accidental offence.

Dataset & Labels

We fine-tuned DistilBERT-base-turkish-cased on the Overfit-GM turkish-toxic-language corpus (see table). At runtime, the five original labels collapse into a binary scheme: if any non-OTHER score exceeds 0.90, the utterance is flagged as PROFANITY (1); otherwise it stays OTHER (0). This strict threshold keeps false positives low while still catching unmistakable toxic speech

Original label	Samples	Binary mapping
OTHER	37663	0
PROFANITY	18252	1
INSULT	10777	1
RACIST	10163	1
SEXIST	945	1

Performance Metrics (Binary Evaluation)

Metric	Score
Accuracy	0.9691
Precision (we ghted)	0.9693
Recall (weighted)	0.9691
F1-Score	0.9691

Methodology

1- ASR – Whisper-large downsamples 48 kHz mic audio to 16 kHz and outputs word-level timestamps.

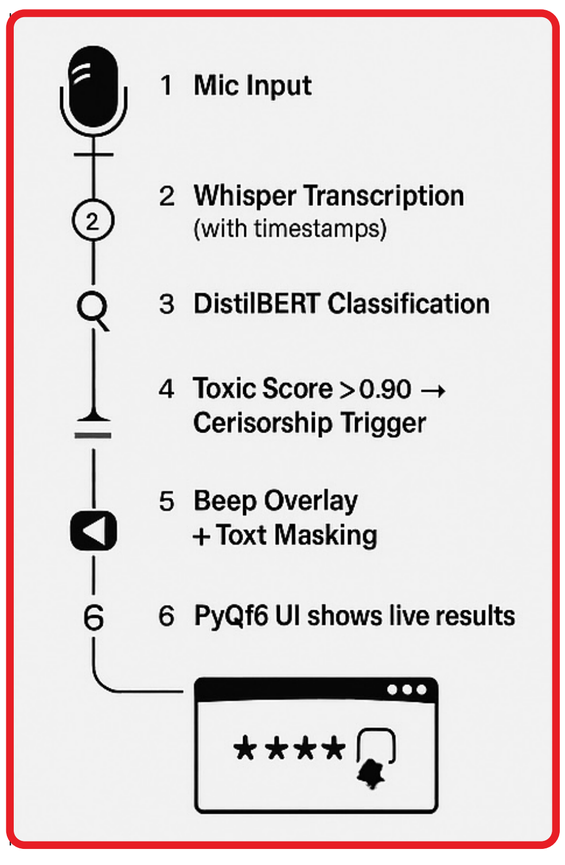
2-Classifier – Each word is scored by our fine-tuned DistilBERT across the five toxicity labels.

3- Logic – Any toxic score > 0.90 triggers censorship.

4- Audio Edit – Flagged spans are merged and overlaid with same-length sine-tone beeps–no timing gaps.

5- UI & Playback – A PyQt6 window shows raw vs. masked text and streams the sanitized audio, all in < 0.4 s on a GTX 1650 Ti.

FLOW CHART



TECHNOLOGY STACK

DistilBERT-Turkish-Offensive – LLM model

Turkish-Toxic-Language – Dataset

Python 3.11 – Language

PyTorch 2.3 – DL framework

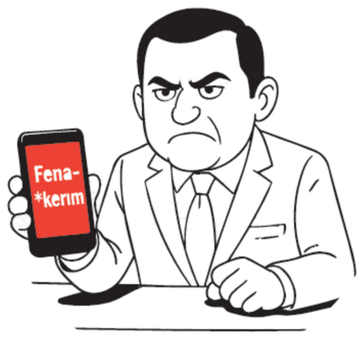
OpenAI Whisper – ASR

Transformers – LLP library

PyDub + FFmpeg – Audio tools

PyQt6 – GUI

NumPy / SciPy – Numerics



Key Advantages

- Privacy by Design. All raw audio stays on the device; nothing is uploaded.
- Word-Level Precision. Whisper’s timestamps make every censor beep align perfectly with the spoken word, avoiding noticeable delays.
- Ambiguity Handling. Words that could be misunderstood as slurs are also masked, lowering the chance of unintended offence.
- Modular Codebase. The ASR model, toxicity classifier, and UI layer are decoupled, allowing easy replacement or extension–e.g., adding English profanity support or streaming directly to OBS.

REAL-WORLD APPLICATIONS

Live TV / radio delay boxes

Classroom lecture capture

Esports & gaming streams

Public kiosks & museums

In Vehicle Information Systems