DigniFy: A Hate Speech Detection Tool

DigniFy is a pioneering technological intervention designed to combat the escalating challenge of online hate speech. Our solution leverages advanced deep learning technologies to create a comprehensive, multi-modal hate speech detection system.

Team Members

Tirath Bhathawala (C165)

Siddhant Uniyal (C154)

Shubham Jaiswar (C153)

Vikas Kewat (C181)

DigniFy

Detect Hate. Prevent Harm. Unite Communities.

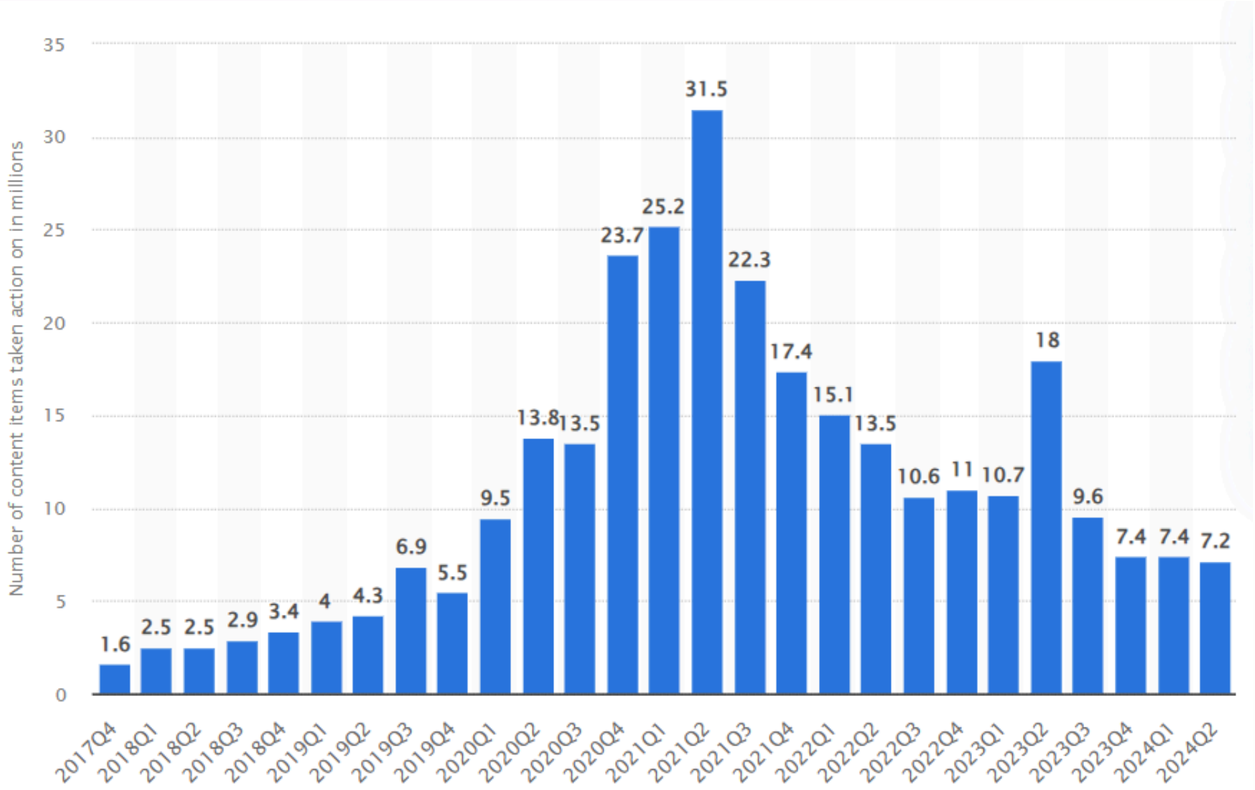
The Urgent Need for DigniFy

Escalating Threat

38% of online users have experienced hate speech. Existing moderation tools fail to capture 60-70% of nuanced hate speech.

Impact

Hate speech can lead to increased psychological distress, reduced online participation for vulnerable groups, and potential escalation to real-world violence.



Quarterly Distribution of Hate Speech Actions on Facebook

Limitations of Current Solutions

Text-Only Analysis

Ignores non-textual communication channels, misses critical contextual nuances, and can't detect sophisticated hate speech encoding.

Language Constraints

Predominantly English-centric solutions with limited support for regional and indigenous languages, resulting in poor performance in multilingual environments.

Detection Accuracy

Typical accuracy rates between 60-75%. High false-positive and false-negative rates, and inability to understand cultural and contextual subtleties.

Potential Use Cases

Corporate Content Moderation

Automated content filtering, implementing inclusive communication policies, protecting brand reputation.

Individual User Protection

Personal content screening, safe browsing recommendations, empowering individual digital experiences.

Community Platform Management

Automated moderation for chat platforms, reducing community toxicity, encouraging constructive dialogue.

Research & Technological Landscape

1 Comprehensive Field Investigation

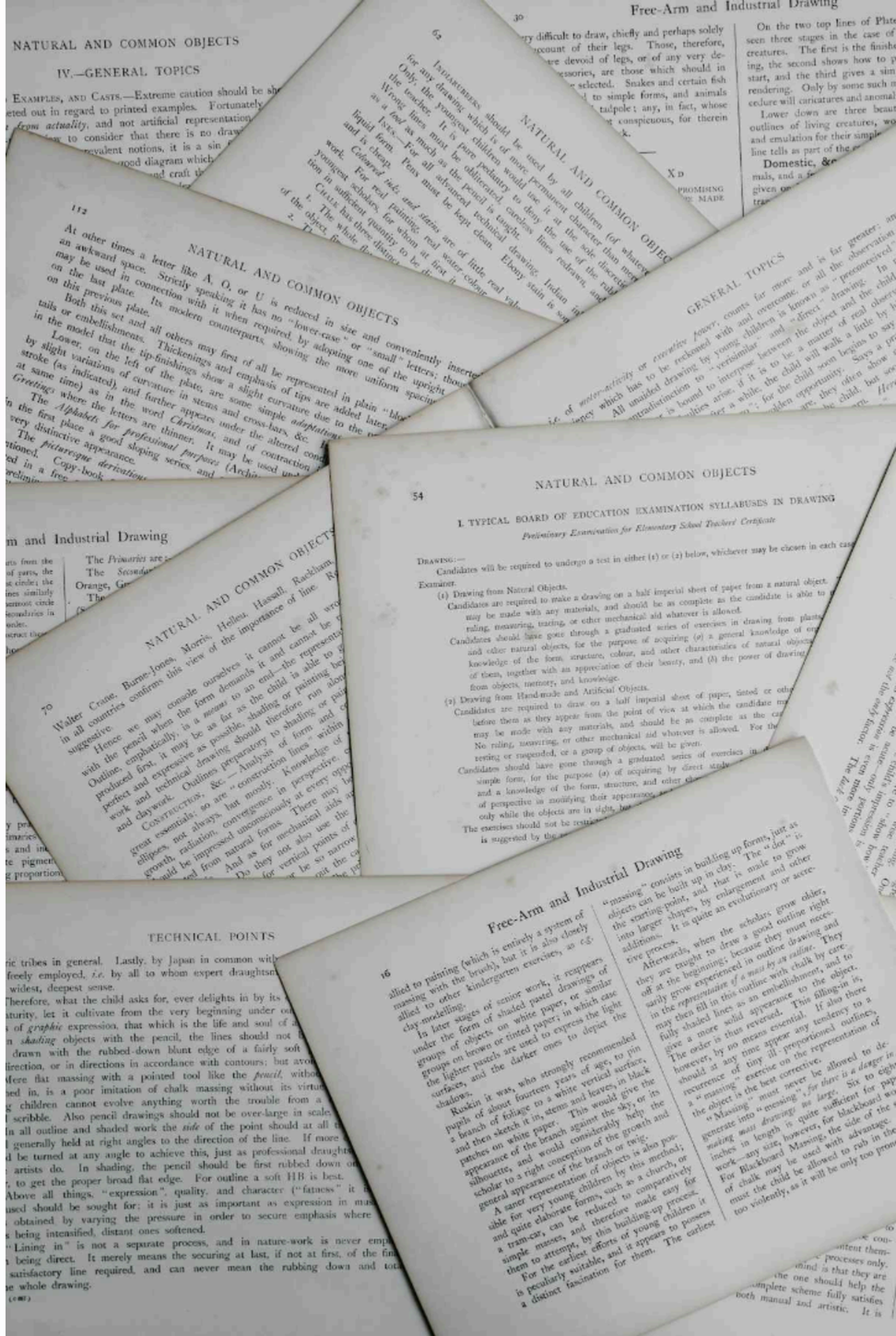
Analyzing online platforms, studying hate speech propagation mechanisms, examining psychological impact studies.

2 Rigorous Literature Survey

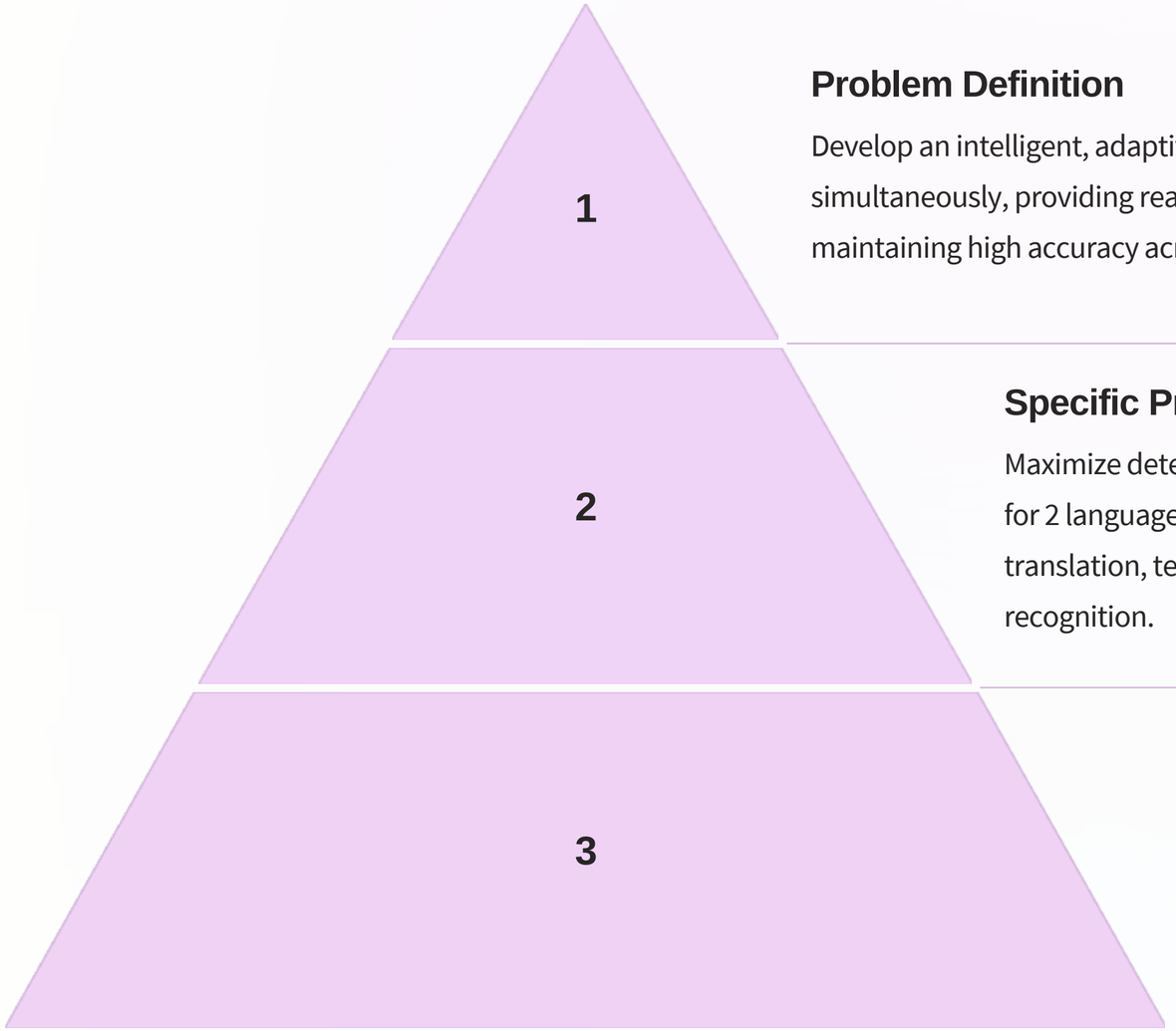
Reviewing key research papers on multilingual and multi-modal hate speech detection, generative AI in hate speech identification, and exploring model performance and limitations.

3 Survey Insights

The need for multi-modal detection capabilities, deep learning-based analysis, comprehensive linguistic and cultural understanding, and real-time processing mechanisms.



Problem Formulation



Problem Definition

Develop an intelligent, adaptive hate speech detection system capable of analyzing multiple communication modalities simultaneously, providing real-time, contextually aware detection, supporting diverse linguistic environments, and maintaining high accuracy across varied content types.

Specific Product Objectives

Maximize detection accuracy, minimal processing time, minimal computational overhead, native support for 2 languages, adaptive learning for linguistic nuances, contextual understanding beyond literal translation, text semantic analysis, voice tone and inflection detection, visual symbol and gesture recognition.

Innovative Technological Approach

First truly multi-modal hate speech detection framework, deep learning with transfer learning capabilities, adaptive machine learning models, comprehensive cross-media analysis.

Experimentation & Results

1

Dataset Composition

Multilingual text collections, diverse audio recordings, culturally representative visual content, ethically sourced and annotated datasets.

2

Preprocessing

NLP preprocessing, holdout method

3

Finetuning

Deep learning hyperparameter tuning , Cross-linguistic validation

4

Results

Successful multi-modal hate speech detection, multi-modal content testing, high accuracy in text and image analysis, promising initial performance across different languages.

Proposed Architectural Design

Text Module

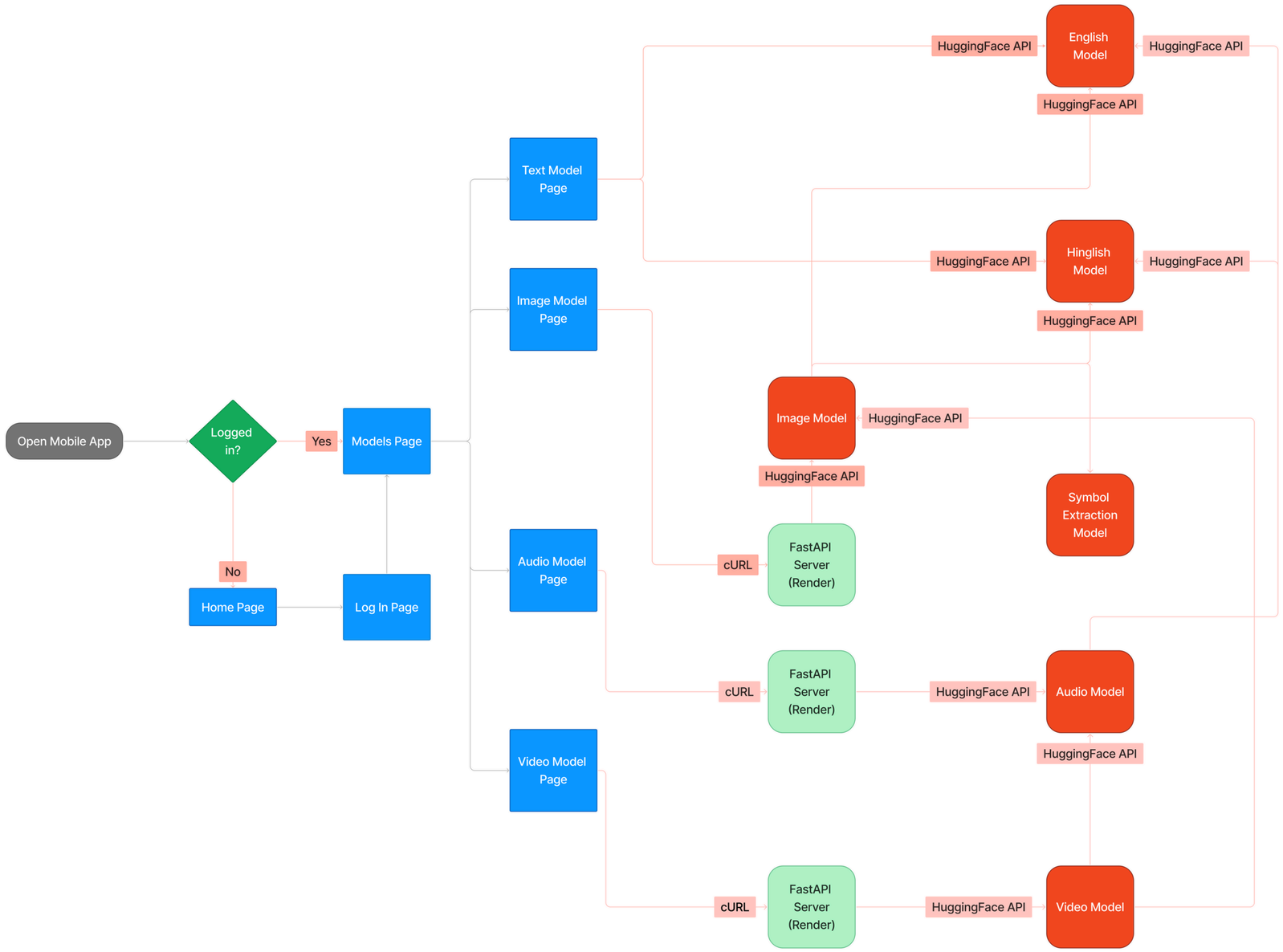
Natural Language Processing (NLP) Engine, semantic understanding beyond literal interpretation, multilingual semantic mapping

Image Module

Computer Vision Intelligence System, symbol and gesture recognition, visual context interpretation

Audio Module

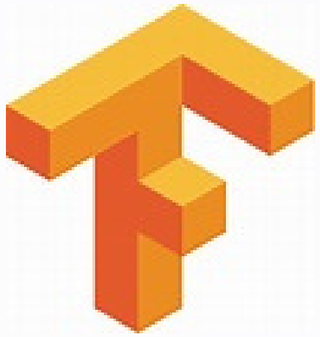
Acoustic Analysis System, Audio Intelligence System, Speech & Sound Recognition



Tech Stack

NumPy

FastAPI



Flutter

1

Deep Learning Frameworks

TensorFlow, PyTorch

2

Natural Language Processing Libraries

NLTK, SpaCy

3

Development Tools

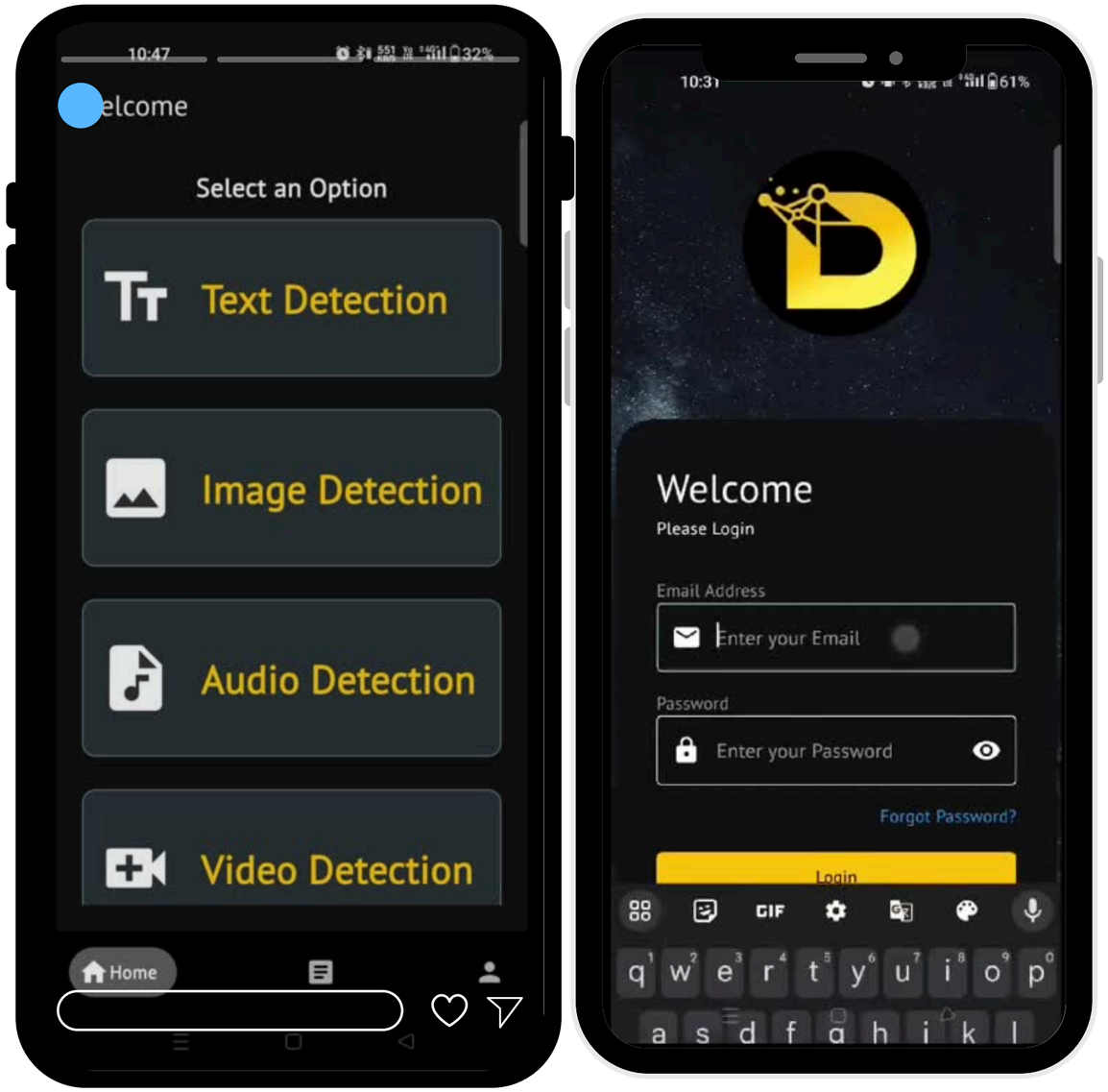
Flutter, FastAPI

Implemetation

- DigniFy is implemented using a multi-layered architecture, leveraging deep learning models for text, image, audio and video.

- We utilize transformer-based models, pre-trained on massive datasets, to capture complex linguistic nuances and contextual cues in hate speech detection.

- Our system undergoes fine-tuning and optimization using a comprehensive dataset curated from various online platforms.



Conclusion: A More Empathetic Digital Future

DigniFy represents a significant step toward creating safer and more inclusive digital spaces. By understanding the complex, nuanced nature of hate speech, we aim to foster environments of mutual respect and understanding.

DigniFy

IPD Team Members

Tirath Bhathawala (C165)

Siddhant Uniyal (C154)

Shubham Jaiswar (C153)

Vikas Kewat (C181)