DigniFy: A hate speech detection tool

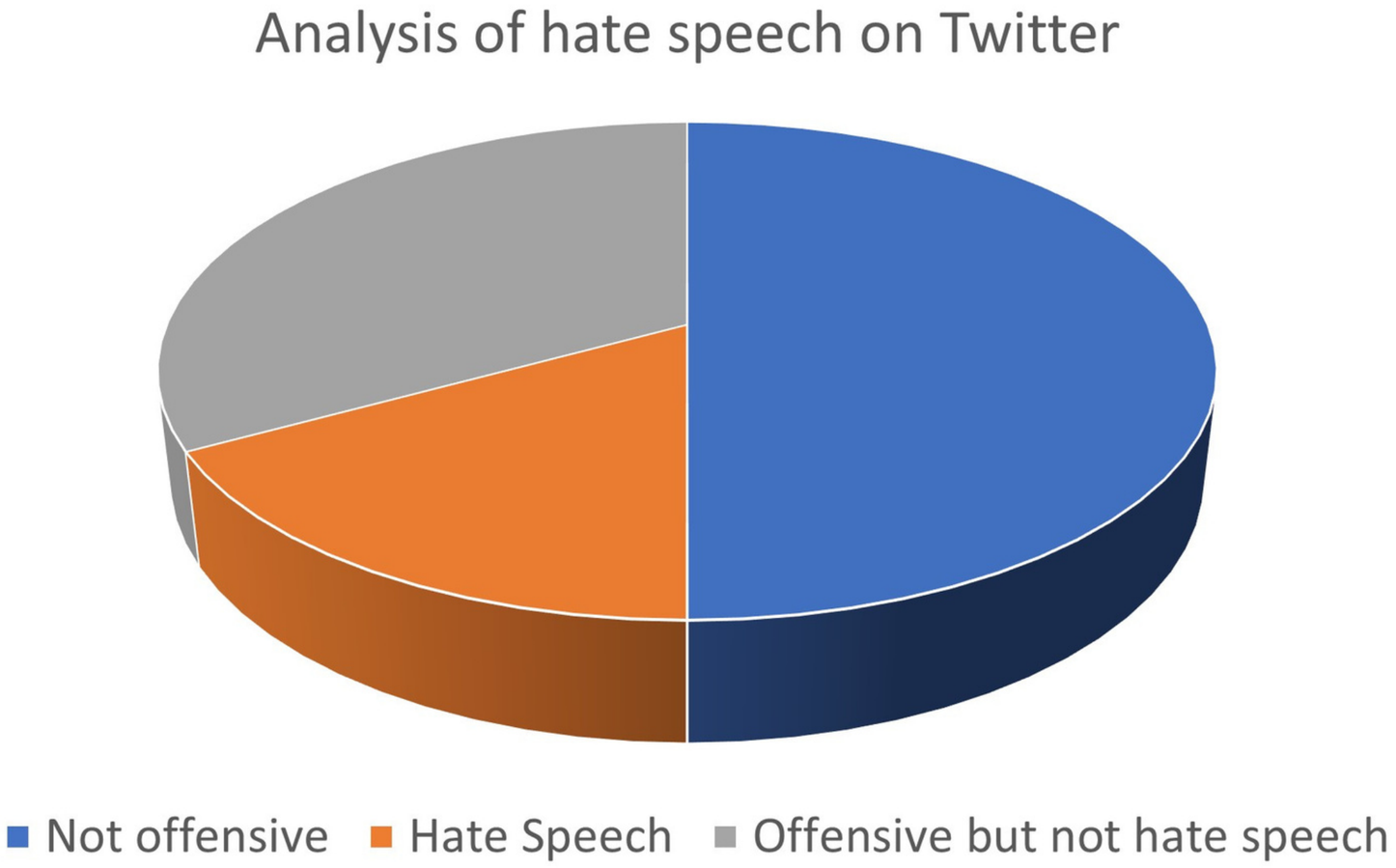
Detect Hate. Prevent Harm.
Unite Communities.

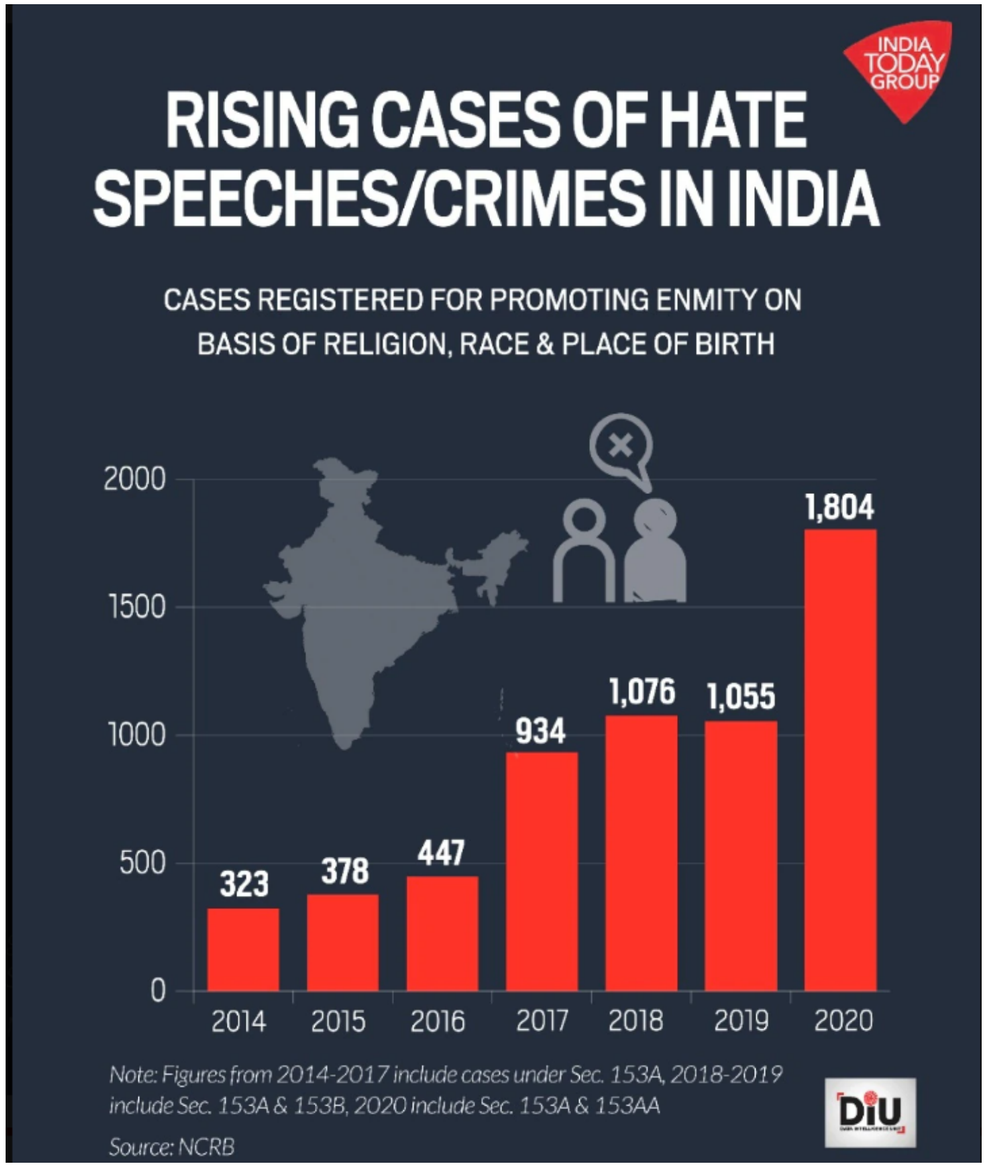
Introduction

- Detecting hate speech is important to protect vulnerable groups and create welcoming online spaces.
- It helps promote empathy, prevent violence, and stop improper language.
- Our proposed product includes a multi-modal multilingual deep-learning based detection tool to find and stop harmful content in various media, making online spaces safer and more inclusive by fighting hate speech.

Significance of the research topic

- Empower digital safety with our hate speech detection powerhouse.





Previous Products:-

- Tilt: Online platform that offers near real-time monitoring to detect hate speech.
- DACHS: Discovers Hate Speech directed at journalists and news outlets, and develops strategies for journalists to counter online hatred.

Drawback:- Only text based classification and corporation focused service.

Reference Research Paper

Title of Research Paper	Finding of the Research Paper	Drawback of the Research Paper	Year	Link
Deep Learning Models for Multilingual Hate Speech Detection	Translation + BERT works best for foreign languages	Does not provide multi-modal support; only takes in multilingual text.	Dec 2020	arXiv
Multi-modal Hate Speech Detection using Machine Learning	Maximum accuracy achieved using Supervised Learning: 87%	Does not use deep-learning based model; relies on SVMs and Naive Bayes model.	June 2023	IEEE Xplore
Generative AI for Hate Speech Detection	Text-davinci was the best model across all datasets , with median f1 score : 0.684	Does not provide multi-modal and multi-lingual support; only takes in text.	Nov 2023	arXiv

Proposed Methodology

Text

Analyses multilingual text to uncover patterns from the given input and classify the text as either hateful or non-hateful.

Voice

Use voice extraction techniques to capture tone and context enhancing ability to identify hate speech in spoken communication.

Image/Video

Algorithm carefully analyze visual content, detecting subtle cues and symbols associated with hate speech.

Methodology (Cont.)

Tool combines text, voice, and image/video analysis to effectively detect and combat hate speech across all digital platforms.

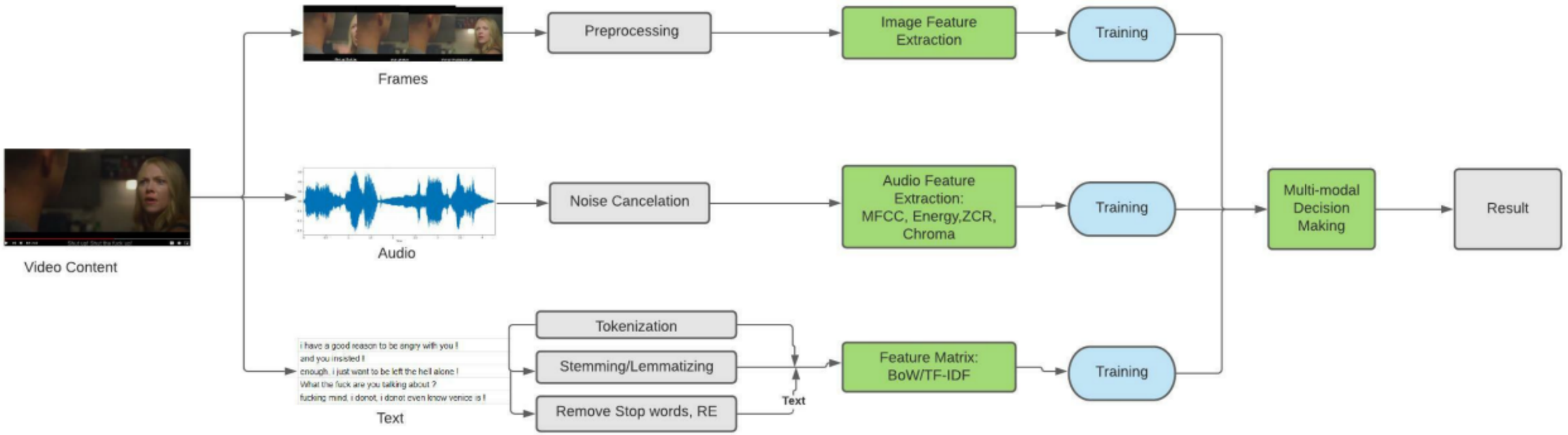


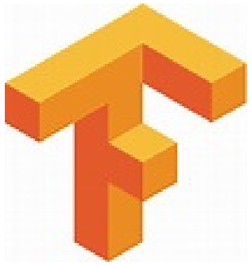
Fig. 1. Methodology for Detecting Hate Speech

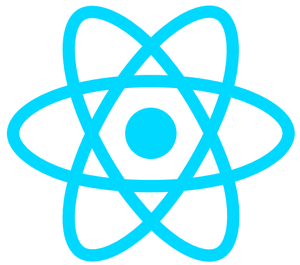
Database Available

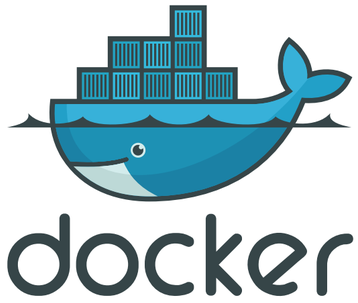
- Text-based and image-based datasets are readily available
- Audio and video datasets will be scraped from YouTube using PyTube

References	Year	Dataset	Dataset Description	Language
[15]	2022	Thomas Davidson, Zeerak Waseem	Datasets divided into categories like "Hate," "Offensive," "Racism," "Sexism," etc.	English
[20]	2021	Offensive Language Identification Dataset (OLID)	Dataset for identifying and categorizing offensive language in social media	English
[29]	2021	Waseem and Hovyf	Dataset of racist and sexist tweets labeled by expert annotators and activists	English
[41]	2021	Stream of-consciousness essays (SoCE) and YouTube (YoTB)	Dataset classified into personality traits like neuroticism, extroversion, etc.	English
[46]	2021	Devanagari Hindi Offensive Tweet (DHOT)	Dataset collected from 150 Hindi-speaking people	Hindi
[56]	2019	Movie reviews, Stanford Sentiment Treebank SST-1, SST-2 datasets	Dataset with labels ranging from very negative to very positive	English

Tech Stacks







HW/SW Requirement

- Hardware: GPUs for accelerated processing to support deep learning algorithms.
- Software: Deep learning frameworks (e.g., TensorFlow, PyTorch) and necessary libraries for natural language processing, audio processing, and image/video processing.

Use case

Company

Companies can give their database and we will filter out hateful comments from it. They can give access to company feedback so that we can enforce a no-tolerance policy.

Individual

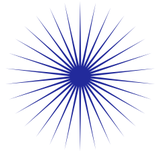
Users can submit documents, images, audio, and videos for hate content detection. If the website contains hate speech, we caution users against visiting it upon clicking the link.

Community

We enhance user experience in Discord/Telegram communities by offering hate content detection bots for improved moderation.

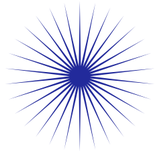
Future prospects

Possible changes to improve DigniFy’s user experience



More language support

Further support for local indigenous languages so that user can filter out hate in any form.



Classification of Hate

Hate can be classified by target (e.g. racism, sexism), intent (individual, systemic), severity (verbal, symbolic, violence), or other factors (motivation, context).

Ban system in chat

If a user comment hateful comment then the user will be flagged and warned. Multiple violations will result in a ban.

Thank You

Vikas Kewat (SAP ID: 60004220181)
Tirath Bhathawala (SAP ID: 60004220101)
Siddhant Uniyal (SAP ID: 60004220202)
Shubham Jaiswar (SAP ID: 60004220112)