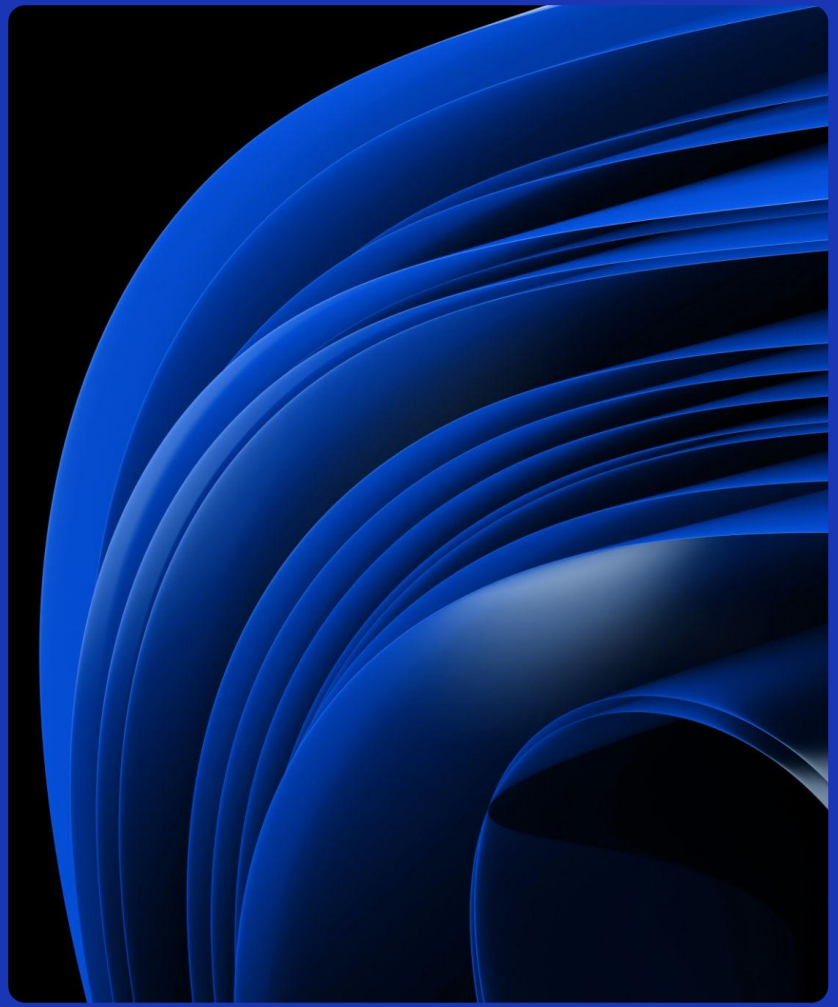
Applied AI Project Proposal

By Karthik, Wei Jun, Jun Ming, Pin Shien
14 January 2025



Problem Statement

Deep dive into the problem at hand

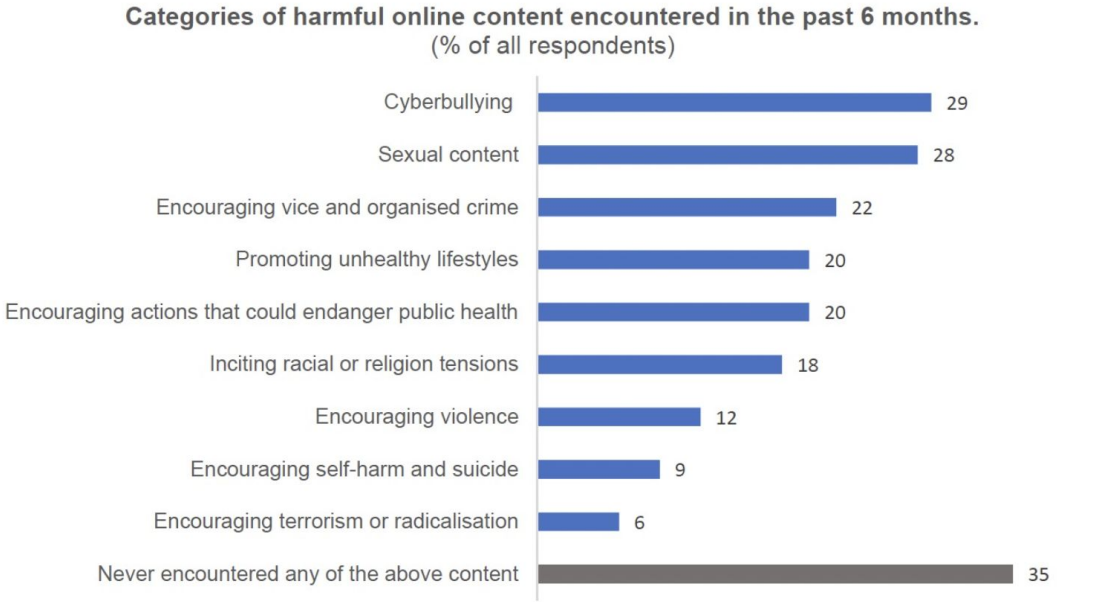
Introduction

- Online trust and safety is a critical global issue, particularly for highly digitalized nations like Singapore, where high internet penetration and a tech-savvy population demand robust governance.

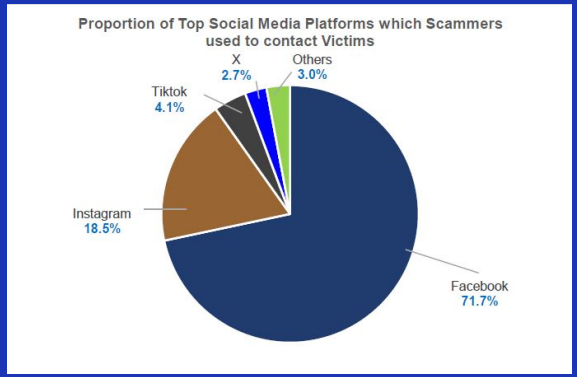
- Existing measures include:
  - Online Safety (Miscellaneous Amendments) Act: Enhances online safety regulations.
  - Centre for Advanced Technologies in Online Safety (CATOS): Focuses on advanced technologies for detecting and countering harmful content.

- However, there are still many issues in detecting online content, especially with more newer types of scams coming out:

Introduction



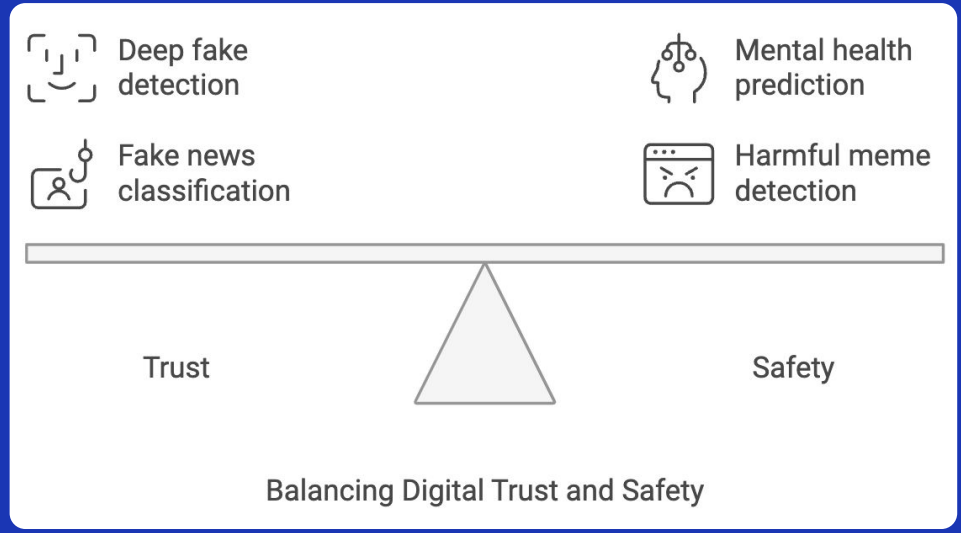
Note:
Examples of "vice and organised crime" include scams, selling illegal goods, and recruitment for criminal activities.



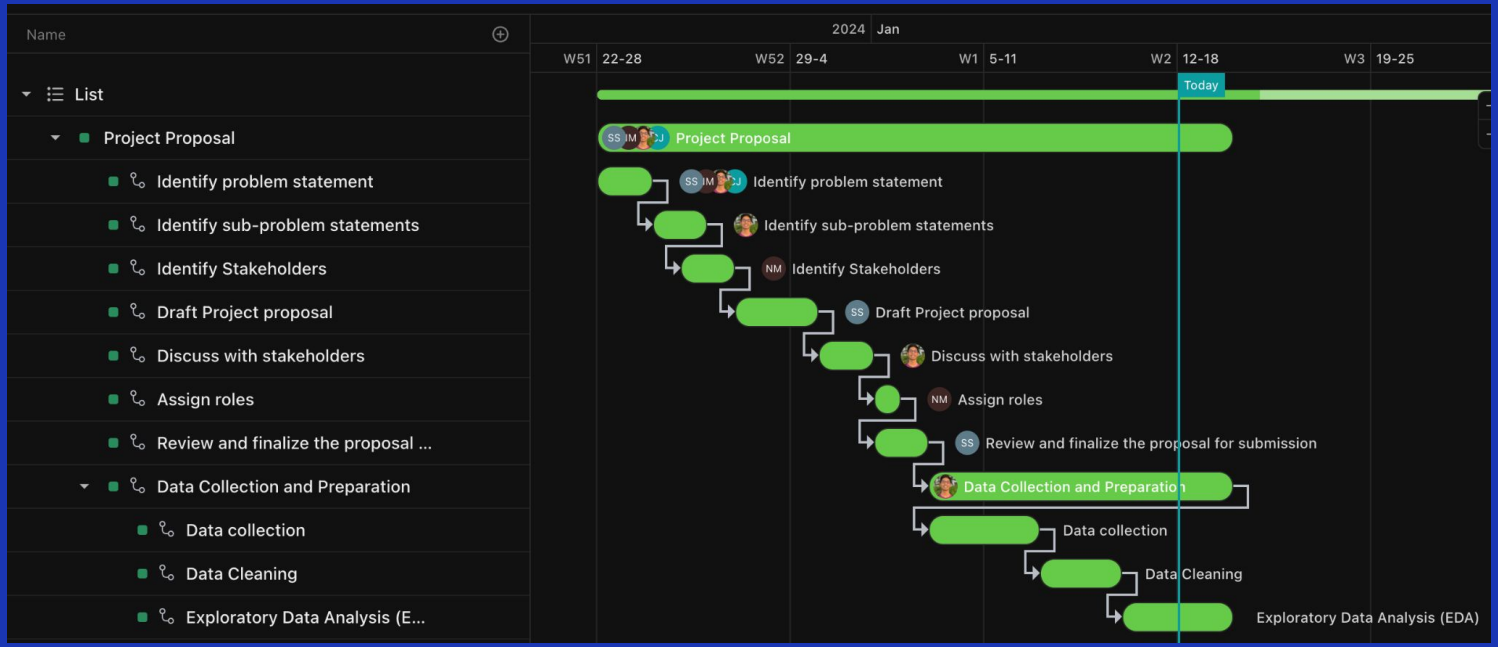
2/3rds
Found harmful content

25%
of incidents reported

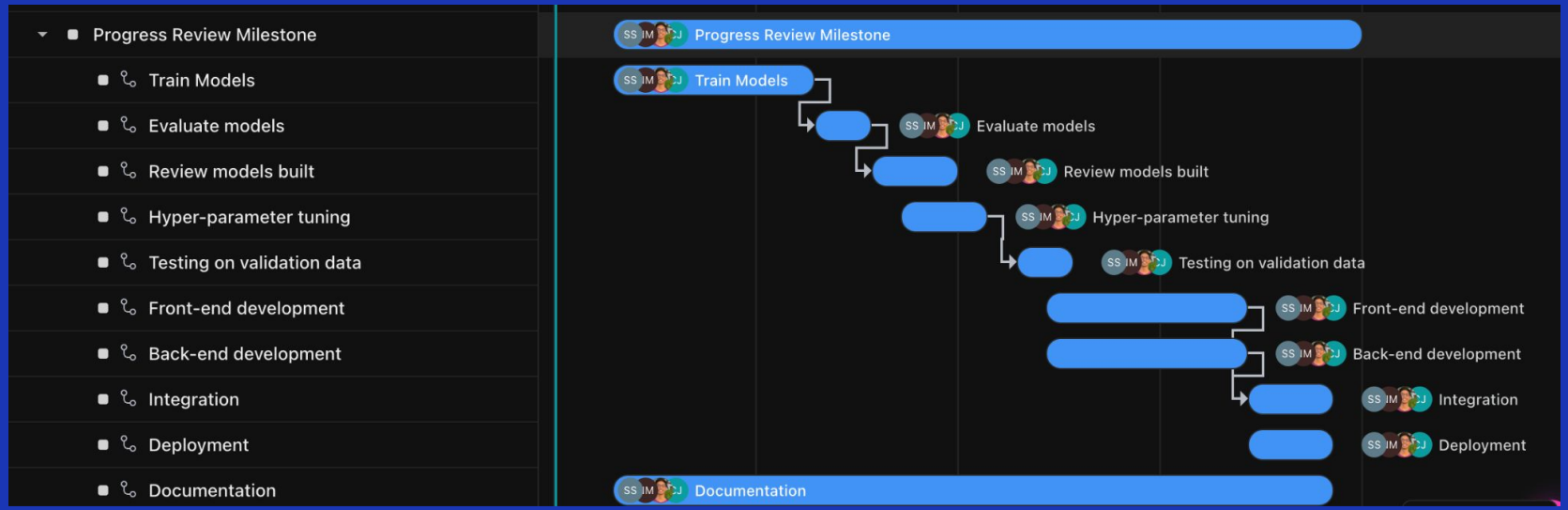
Solution



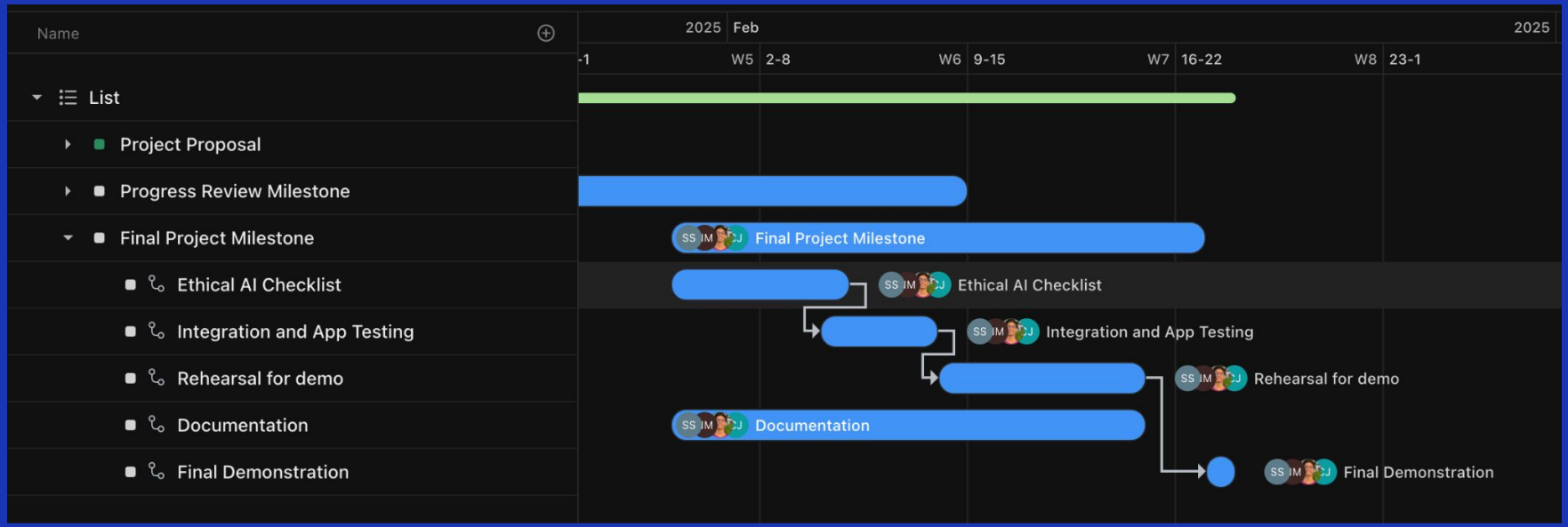
Timeline - Phase 1



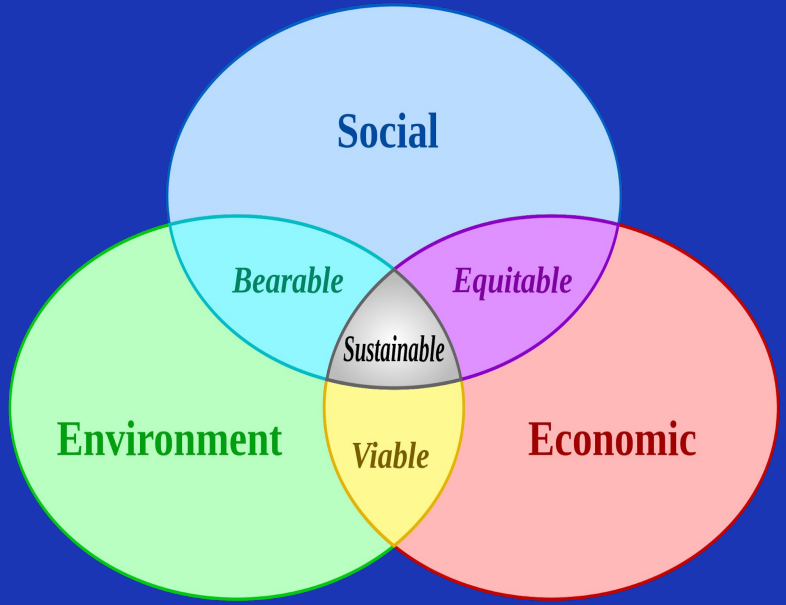
Timeline - Phase 2



Timeline - Phase 3



Impact - sustainability



- Focus on the least looked at area: social sustainability
  - Frequently overshadowed by environmental concerns
- Foster online spaces where individuals feel respected, protected, and empowered to engage without fear

Impact - sustainability

Ensure safer online spaces by mitigating cybersecurity threats, removing harmful content and combating misinformation.

Promotes inclusive digital communities with personalized safety settings, contributing to secure and sustainable urban living

Target Audience

Stakeholders

Integrate the models into social media platforms to enhance safety, trust and reporting

Target Audience

Users of social media platforms

Introducing

 Poseidon

Surf the Web 🌊 with Safety and Trust 🔒

Hateful Meme Classification

By Karthik

Problem Statement

- Memes have become a dominant form of online communication.
- Malicious actors exploit memes to disseminate hateful content, targeting racial, religious, and political sentiments under the guise of humor.
- Studies estimate that 60% of memes contain divisive or harmful content
- Millennials consume 20–30 memes daily, increasing the risk of exposure to harmful material.

The aim to reduce the percentage of Singaporeans encountering harmful online content from the current 74% (as of 2024) to a significantly lower number.

Good Meme

Bad Meme

Pain Points

Anxiety, Stress and Insecurity

Exposure to undetected offensive memes can exacerbate psychological distress, leading to increased anxiety, emotional vulnerability, and feelings of insecurity among social media users

Impact Reputation

Damage the reputation of social media platforms and person/org posting the meme, as they may be seen as facilitating the spread of harmful content and being insensitive

Lead to Violence

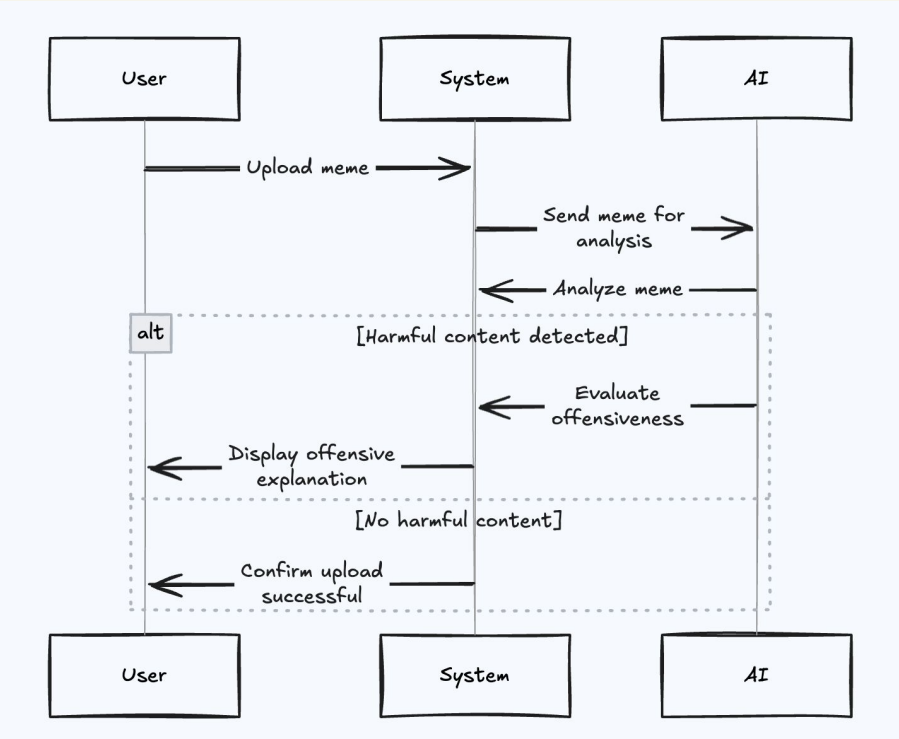
Normalize extremist behaviors and ideologies, potentially influencing real-world events and even inciting physical violence

Examples

2018: Public outrage over students posting 9/11 memes

2022: Criticism from gender advocacy groups

Solution



- **Key Issues:** *Challenging to interpret both image and text together (one could be benign and the other can be harmful)*
- A multimodal classifier that detects if a meme is harmful or not.
- If it is harmful, explain to the user why it is.
- Human in the loop

Objectives Identified

Data Visualization

- Cleaning both textual and image datasets
- Ensure they are clean and error free

Experimentation with Models

- Consider SOTA embedding models and different neural network architectures
- Experiment with large range of hyperparameters

Performance Benchmarks

- Use AUROC as main metric as it shows the discriminative ability of the model between 2 classes.
- Widely used in multimodal classification papers and FHMC competition

Additional Features

- Adding explainbility through VLM to determine which part of the meme is offensive
- Give suggestions on how to make it less offensive

Datasets Used

facebook

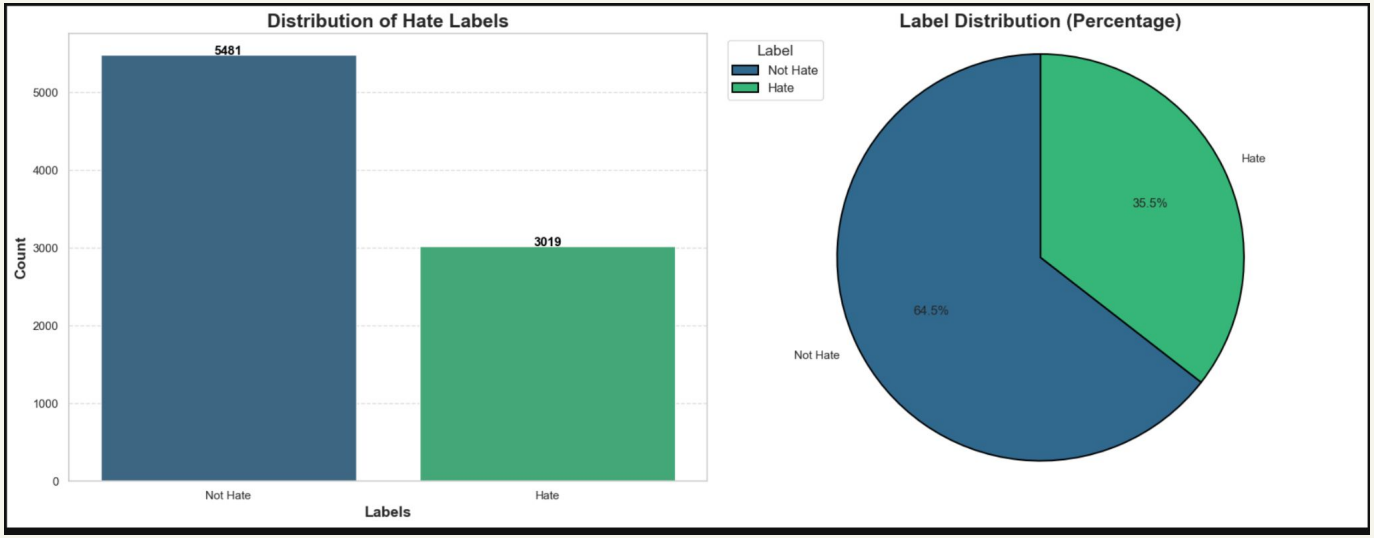
FACEBOOK HATEFUL MEME DATASET

- **Data Quality:** Relatively Clean
- **Columns:** ID, Image, Text, Label
- **Rows:** ~10K Rows
- **Source:** Kaggle Datasets

WEB SCRAPED HARMFUL MEME DATASET

- **Data Quality:** Unclean Data
- **Columns:** None (Folder of Images)
- **Source:** Bing Search (100+ offensive meme topics queried)

Data Imbalance



- Large number of non-hate memes as compared to hate meme, leading to class imbalance.
  - Additional memes will be web scraped from the internet to balance the hate class

Web Scraping - Data Collection

```
queries = [
    "racism", "sexism", "homophobia", "transphobia", "hate speech", "violence",
    "bullying", "body shaming", "mental illness stigma", "discrimination",
    "hate crime", "toxic masculinity", "misogyny", "misandry", "anti-Semitism", "xenophobia", "ageism",
"fatphobia",
    "nazi", "white supremacy", "KKK", "neo-nazi", "terrorism",
    "gun violence", "school shootings", "abortion debate", "anti-vaccine", "conspiracy theories",
    "covid misinformation", "climate change denial", "fake news", "radicalization", "ISIS",
    "extremism", "terrorist attacks", "hate speech speech", "slut-shaming",
     "domestic violence", "racist humor", "sexist jokes",
    "anti-LGBTQ", "transgender jokes", "homophobic jokes", "feminist hate",
    "bullying memes", "stereotypes", "meme harassment", "dark humor", "graphic violence",
    "gore", "death threats", "self-harm", "drug abuse",
```

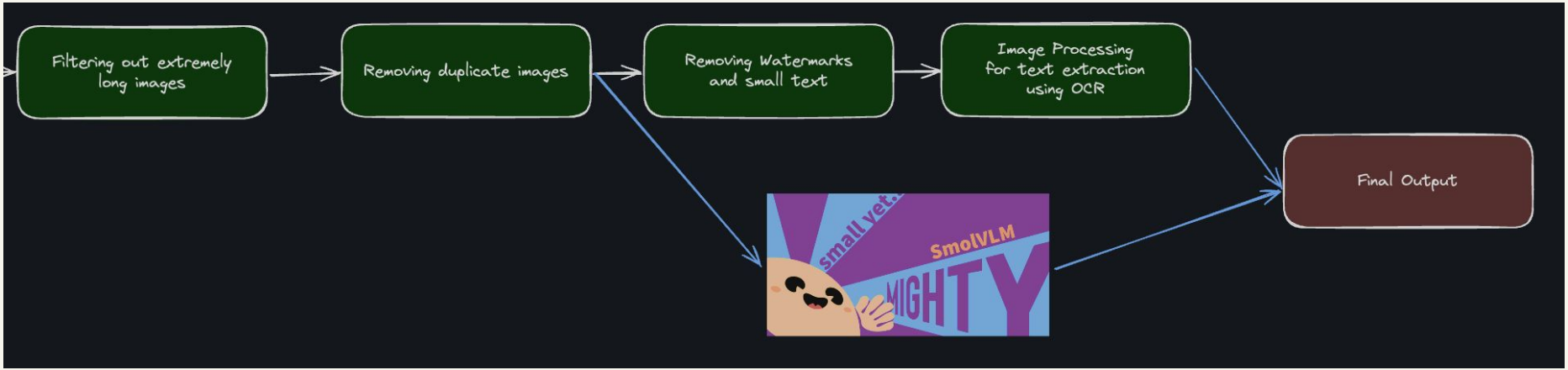
```
for query in tqdm(queries, desc="downloading memes from query", unit="query"):
    if "meme" not in query:
        query += " meme offensive"

    downloader.download(query, limit=20,  output_dir='dataset', adult_filter_off=True, force_replace=False, timeout=60,
verbose=True)
```

```
================================================

[%] Downloading Image #1 from https://www.gannett-cdn.com/presto/2020/06/13/PSIF/32f21d6b-c709-48df-807f-7f4d5eb3b431-Min
nesota_Ave_Protest_009.JPG?crop=5519,3105,x0,y280&amp;width=3200&amp;height=1801&amp;format=pjpg&amp;auto=webp
[!] Issue getting: https://www.gannett-cdn.com/presto/2020/06/13/PSIF/32f21d6b-c709-48df-807f-7f4d5eb3b431-Minnesota_Ave_
Protest_009.JPG?crop=5519,3105,x0,y280&amp;width=3200&amp;height=1801&amp;format=pjpg&amp;auto=webp
[!] Error:: HTTP Error 406: Not Acceptable
[%] Downloading Image #1 from https://ichef.bbci.co.uk/news/1024/branded_news/7924/production/_90821013_tweet_976.jpg
[%] File Downloaded !

[%] Downloading Image #2 from https://ichef.bbci.co.uk/news/976/cpsprodpb/A277/production/_90819514_newtweet_976.jpg
[%] File Downloaded !

[%] Downloading Image #3 from https://ichef.bbci.co.uk/news/800/cpsprodpb/7924/production/_90821013_tweet_976.jpg
[%] File Downloaded !

[%] Downloading Image #4 from https://pics.conservativememes.com/want-to-end-racism-stop-telling-them-theyre-racist-turni
n-31911633.png
```

Web Scraping

- List of 100+ offensive meme topics were generated using chatGPT and queries using the bing image downloader API
- Top 20 results from each topic were queried
- 2637 images were downloaded

Web Scraping - Data Transformation



Web Scraping - OCR comparison

```
Extracted Text:
HOW MANY CONSPIRACY
THEORIES DOES
A PERSON HAVE TO 6e
BELIEVE BEFORE THEY ARE
RECOGNIZED AS HAVING `
MENTAL HEALTH PROBLEMS?
IS IT ONE MORE THAN THE NUMBER YOU BELIEVE?
```

```
% 4 ~ HOW MANY CONSPIRACY
owe THEORIES DOES ¥
% APERSON HAVE TO #=~
wf _ BELIEVE BEFORE THEY ARE
4 .#) RECOGNIZED AS HAVING *
” MENTAL HEALTH PROBLEMS?

“kb

1S IT ONE MORE THAN THE NUMBER YOU BELIEVE?
```

FINAL OUTPUT

HOW MANY CONSPIRACY THEORIES DOES a of PERSON HAVE of of BELIEVE BEFORE THEY ARE RECOGNIZE a of HAVING a MENTAL HEALTH PROBLEMS a a a of ONE MORE THAN THE NUMBER YOU BELIEVE

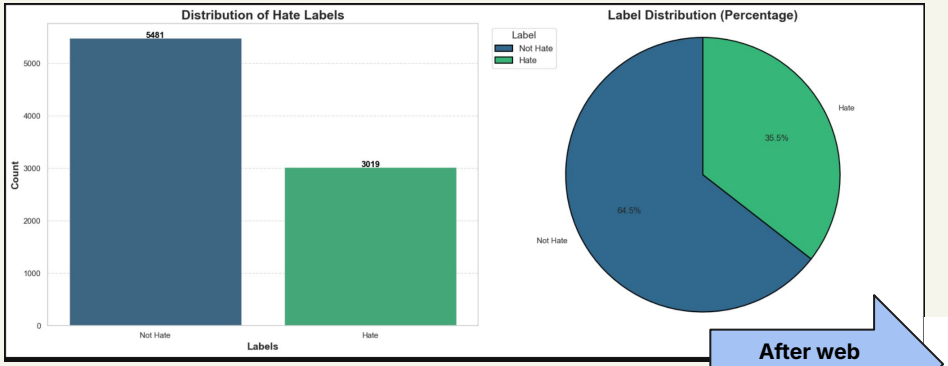
Tesseract OCR

VLM-OCR Sample Output

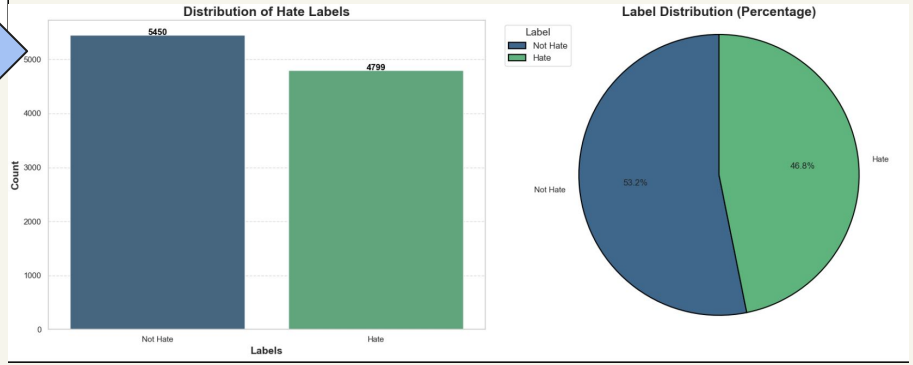
```
Result for Image 1: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: Stay in drugs, eat your school and don't do vegetables.
Result for Image 2: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: What kids think communism is.
Result for Image 3: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: Did you know all war crimes are legal only if you win.
Result for Image 4: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: This is the homophobe flag.
Result for Image 5: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: People who judge you for the colour of your skin...aliens!
```


```
User:<image>Extract all the meme text from the provided image.
Assistant: 不知道什么鬼话，他们只是在说话，我们只是在看。
```

Exploratory Data Analysis - General



After web scraping

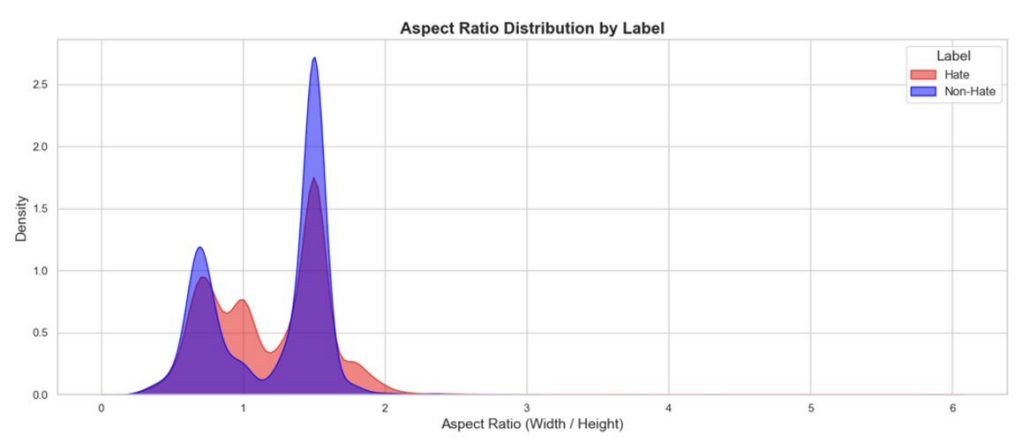


Exploratory Data Analysis - Text

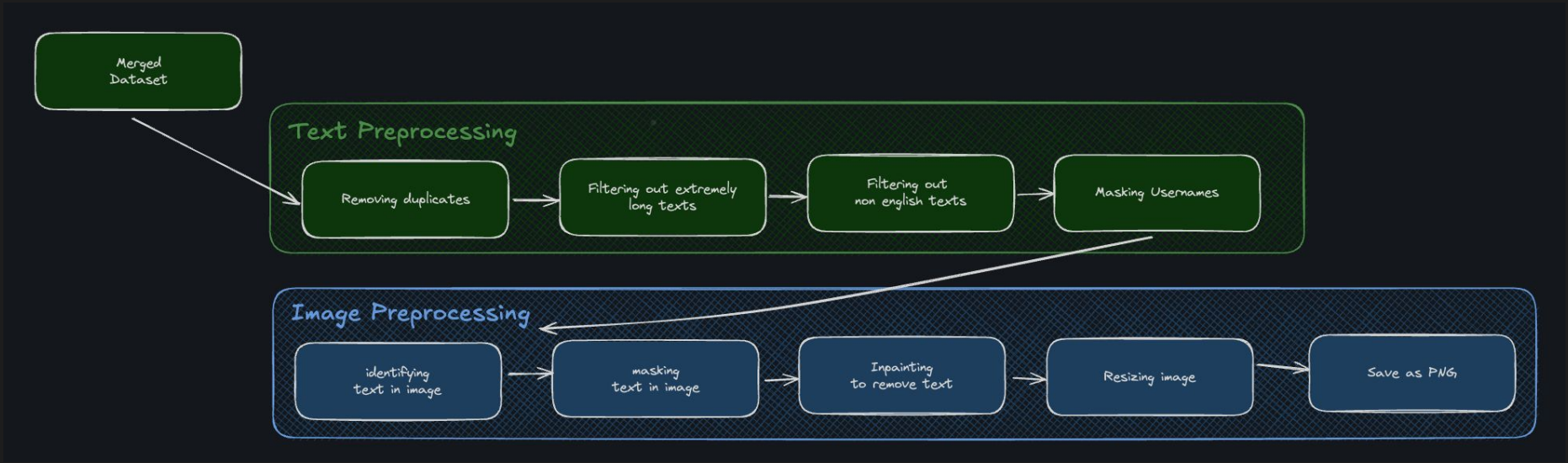
Title	Insights	Data Preparation Solution
Identification of outlier and non-english texts	• Non english texts around 4% of total dataset • Outlier texts have (> 100 words) and (> 50 characters per word)	• Remove these records from the analysis as they are not indicative of actual memes
Identification of POS tags, NER tags, and frequent words	• Hate speech contains more financial references and political/religious entities • Non hate speech contains more ! marks than normal	
Identification of websites, usernames and punctuations	• Usernames and websites were present in the data. • Websites were critical to understanding of meme, so it was retained.	• Mask the usernames in the text • Don't mask the names as they have sentiment.

Exploratory Data Analysis - Image

Title	Description
Aspect Ratio Analysis by target label	· The aspect ratio was plotted as a kernel density estimate plot and split across the target label. · Hateful memes often have a square aspect ratio (1:1), showing a pronounced peak at this ratio, though both hate, and non-hate memes exhibit a bimodal distribution with overlap.



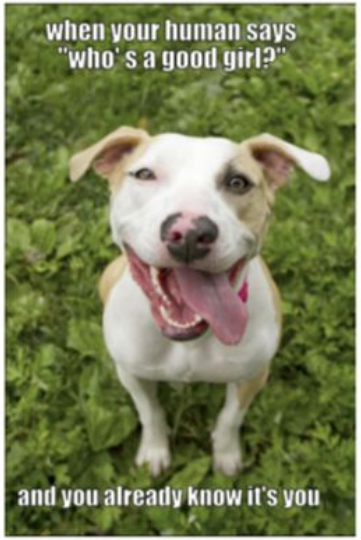
Data Pipeline



- For the embeddings, it will be considered when building the model.
- Looking to experiment with different fusion techniques such as early and late fusion

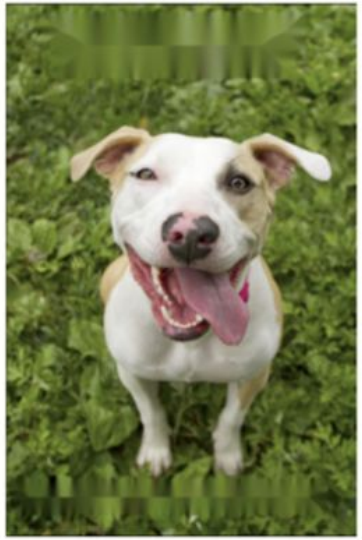
Data Preparation - Image Processing

Original Image



Binary Mask

Text Removed



Data Preparation - Image Processing

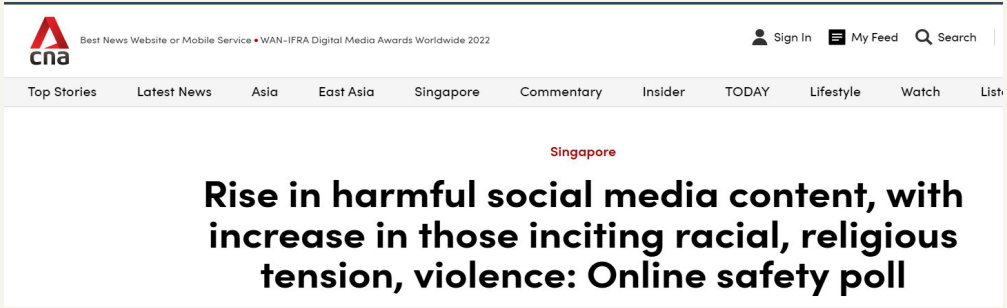
Image resizing using padding + center crop
(512 × 512 image size)

Leveraging Social Media Text for Mental Health Insights

By Pin Shien

Problem Statement

- Social media has become a pervasive part of daily life, offering numerous benefits such as connectivity and information sharing
- However, it has also introduced significant challenges, including the rise of harmful content that adversely affects mental health
- A government survey in Singapore revealed that 74% of respondents encountered harmful content online, a sharp increase from 65% in the previous year
- Despite these alarming trends, action against such content remains limited. 60% of individuals chose to ignore such content, leaving its potential negative impact unaddressed



Pain Points

Algorithmic Amplification

Social media algorithms often prioritize engagement, inadvertently promoting sensational or harmful content that can negatively affect mental health.

Social Pressure and Fear of Missing Out (FOMO)

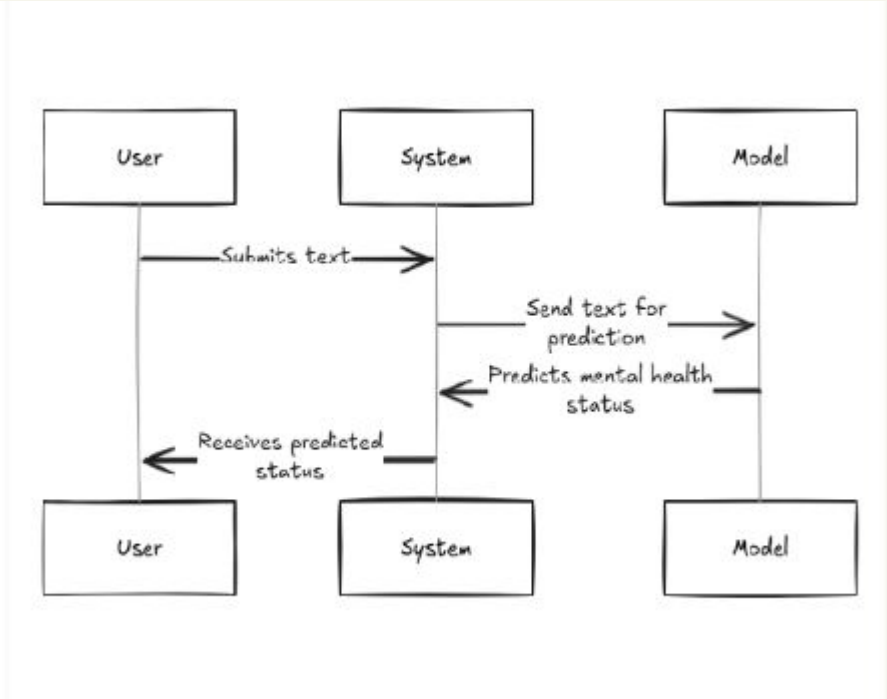
Many users feel compelled to remain active on platforms despite encountering harmful content, fearing disconnection from their social circles or trending discussions.

Youth Vulnerability

Younger users, who often lack the coping mechanisms of adults, are particularly vulnerable to the mental health effects of harmful content, increasing risks of long-term psychological issues.

Proposed Solution

- Build a text classification model to analyze social media posts (e.g., Twitter, Reddit) and categorize them into mental health labels (e.g., Normal, Depression, Anxiety, Stress, Suicidal).
- Identifies at-risk individuals based on their posts for timely intervention.
- Raises awareness and reduces social stigma through data-driven findings.



Objectives Identified

Data Visualization

- Visualize and prepare the data, perform EDA to analyze class distributions and detect anomalies.

Experimentation with Models

- Test embedding methods, machine learning, and deep learning models with hyperparameter tuning.

Performance Benchmarks

- Use F1-score, Precision, Recall etc to measure model effectiveness.

Additional Features

- Apply explainability techniques like SHAP for transparency and fine-tune pretrained models for accuracy and contextual understanding.

Dataset

- This dataset is a curated collection of mental health statuses, tagged from various statements, compiled from multiple sources for sentiment analysis.

Features:

- **unique_id**: A unique identifier for each entry.
- **Statement**: The textual data or post.
- **Mental Health Status**: The tagged mental health status of the statement.

The dataset consists of statements tagged with one of the following seven mental health statuses:

- **Normal**
- **Depression**
- **Suicidal**
- **Anxiety**
- **Stress**
- **Bi-Polar**
- **Personality Disorder**

Data

Steps	Details	Rationale
Identify & Remove Missing Data	Found 362 missing values in the statement column. Missing values removed to avoid biases in analysis.	Removing incomplete data ensures that models rely on meaningful and complete inputs.
Preprocessing Steps	- Convert to lowercase - Remove extra spaces - Remove punctuation and numbers - Tokenize words - Remove stopwords - Lemmatize with POS tagging	The steps clean and standardize text data for analysis by removing noise and transforming it into a structured format.

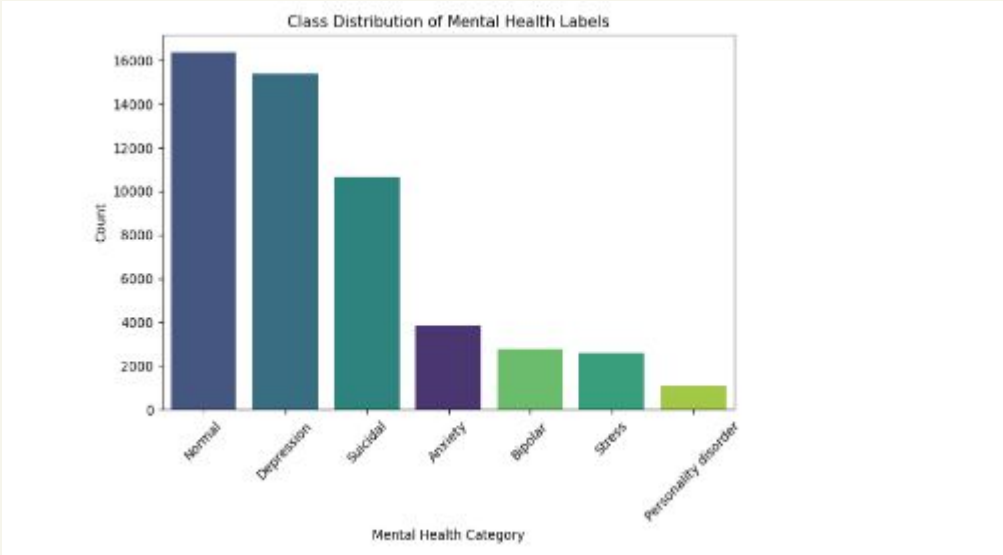
Word Cloud

Insights:

- The word "feel" is the most frequent across all mental health statuses
- Individuals primarily express their mental health experiences through their emotions.
- Words like "go," "work," "think," and "want" frequently appear, reflecting actions, routines, and desires that intersect with mental health.

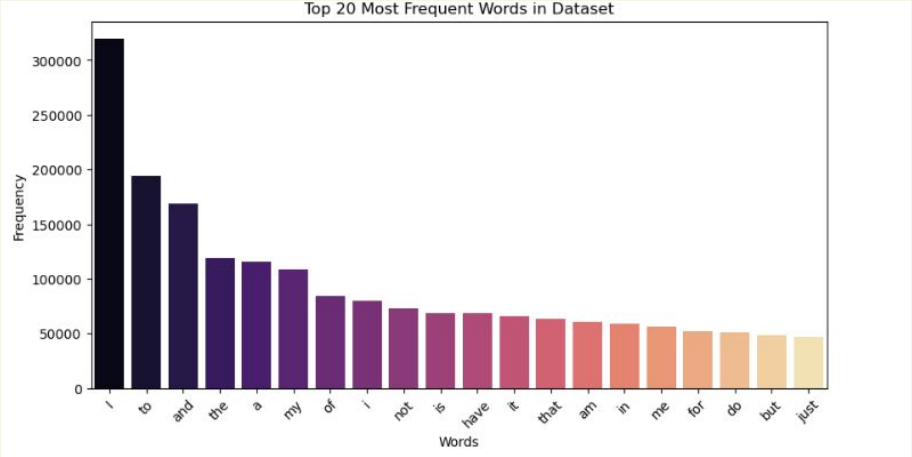
Data Prep / EDA

Class Imbalance: There is a noticeable class imbalance, with the "Normal" and "Depression" categories dominating the dataset. This imbalance could impact model training and may require techniques such as class weighting to address it.

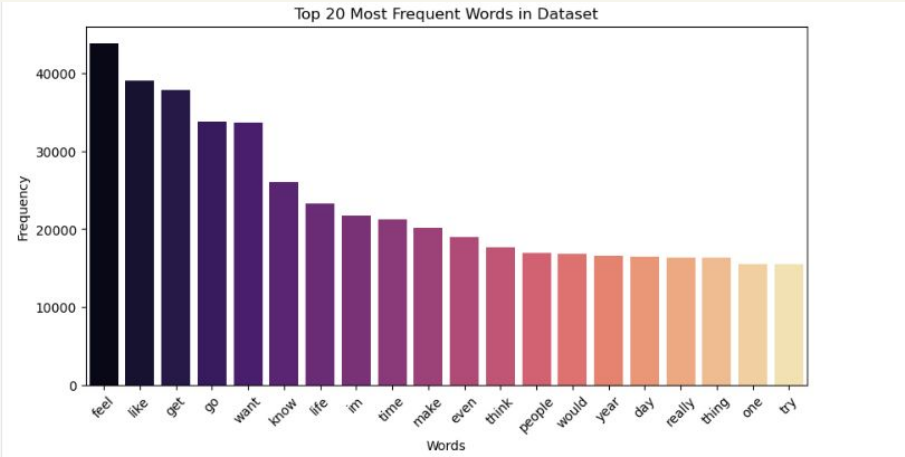


Data Prep/ EDA

- Before Removing Stopwords/Lemmatization



- After Removing Stopwords/Lemmatization



- **Shift to Meaningful Words**: The top 20 words now primarily consist of meaningful and contextually significant terms, reflecting emotions, actions, and thoughts (e.g., "feel," "like," "life," "want," "time").

Detecting Deep Fake Misinformation

By Wei Jun

Problem Statement

Addressing the Rise of Deepfake Misinformation with AI-Powered Detection

- Deep fake of Lee Hsien Loong promoting investment scam on social media.
- Realistic fake videos are often used for scams, spreading false information, and damaging reputations
- Critical need for a robust and reliable deepfake classification model that can accurately identify manipulated videos, helping to prevent misinformation and protect public trust.

Pain Points

Erosion of Trust

Erodes trust in media and public figures by making it harder to tell real from fake content.

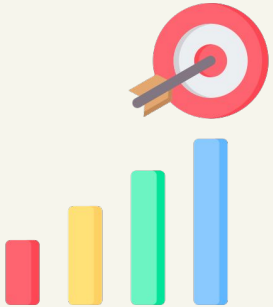
Public Safety Concerns

Manipulate public opinion, incite violence, and spread false information, causing social and political instability.

Increased Vulnerability to Scams

People are more vulnerable to scams like voice phishing and identity theft, as fake voices and videos are harder to detect.

Goals & Stats



Goals

- Prevent online users from falling prey to deep fake scams
- Create a safer and more trustworthy space on social media
- Promote understanding of deepfakes and how to spot them

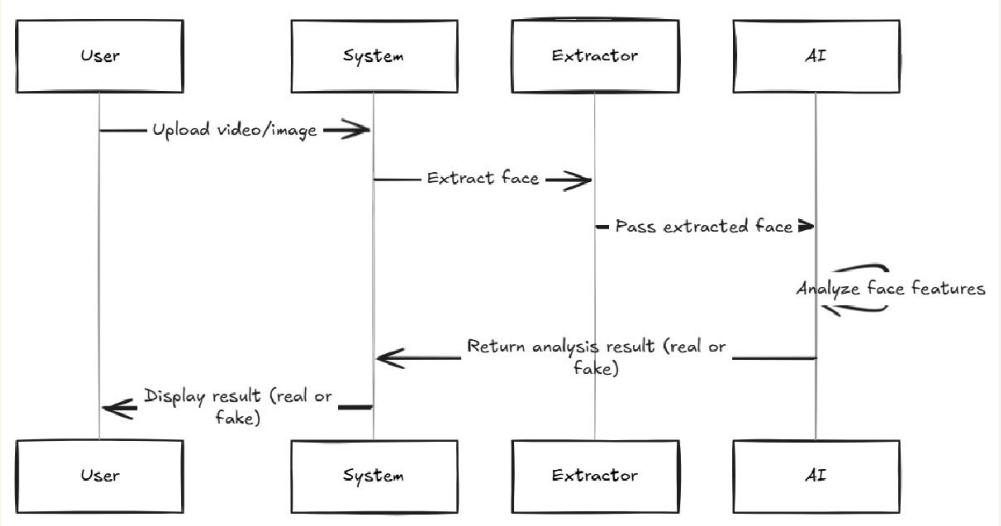
Statistics

- 70% of people said they aren't confident that they can tell the difference between a real and cloned voice. [9]
- Deep Fake fraud increased by 1,740 percent in North America and by 1,530 percent in the Asia-Pacific region in 2022. [9]
- DeepFaceLab claims that more than 95 percent of deepfake videos are created with its open-source software [9]

Solution

- An image classifier that identifies if a video is a deepfake.
- Lightweight for fast identification (10s)
- Reliable performance for trustworthiness
- Human in the loop, due to high severity

Key issues: Factors like lighting, compression, noise, and resolution can obscure deepfake artifacts.



Objectives Identified

Data Visualization

- Perform **exploratory data analysis (EDA)** to understand image dataset characteristics.
- Clean and balance the data

Experimentation with Models

- Utilize Convolutional Neural Networks (CNNs) and pre-trained models with transfer learning.
- Perform hyperparameter tuning to enhance performance and prevent overfitting.

Performance Benchmarks

- Prioritize Recall (0.75) to minimize missed deep fakes and reduce misinformation risks.
- Aim for an F1-Score of 0.7 as a balanced measure of precision and recall.

Additional Features

- Integrate **Google Face Detection API** for face extraction from videos/images
- Pass the cropped faces to the deep learning model for deep fake detection.

Data

Kaggle dataset

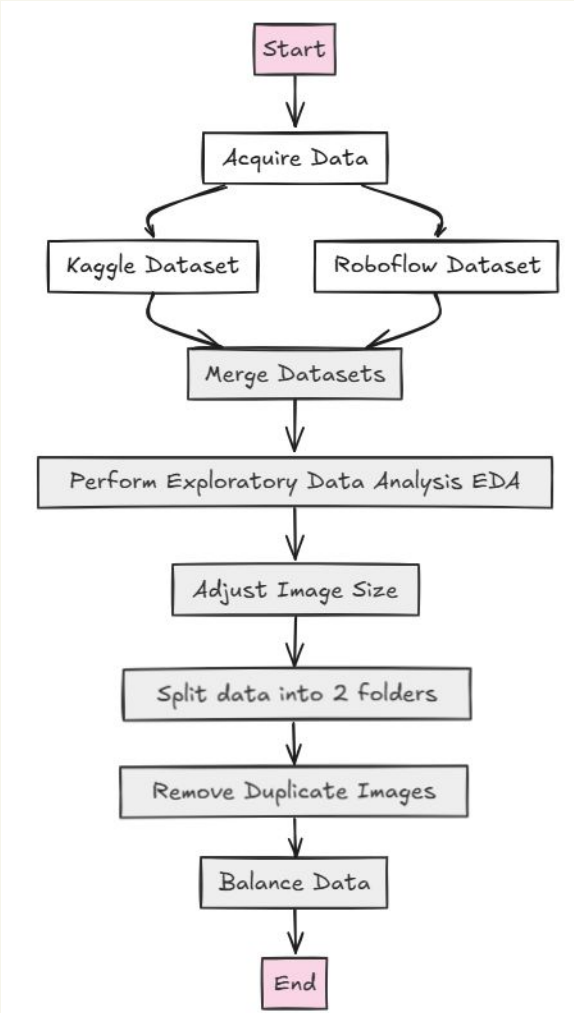
- Subset of this dataset, which has over 470GB worth of videos of real and deep fake videos.
- Over 95,000 images of faces extracted from the first frame of each video.

Roboflow dataset

- Its images have filters applied, (i.e 1 image has 4 versions with different filters)
- Helps the model to be more robust.
- Contains close to 30,000 images

Undersampling was used to balance the data

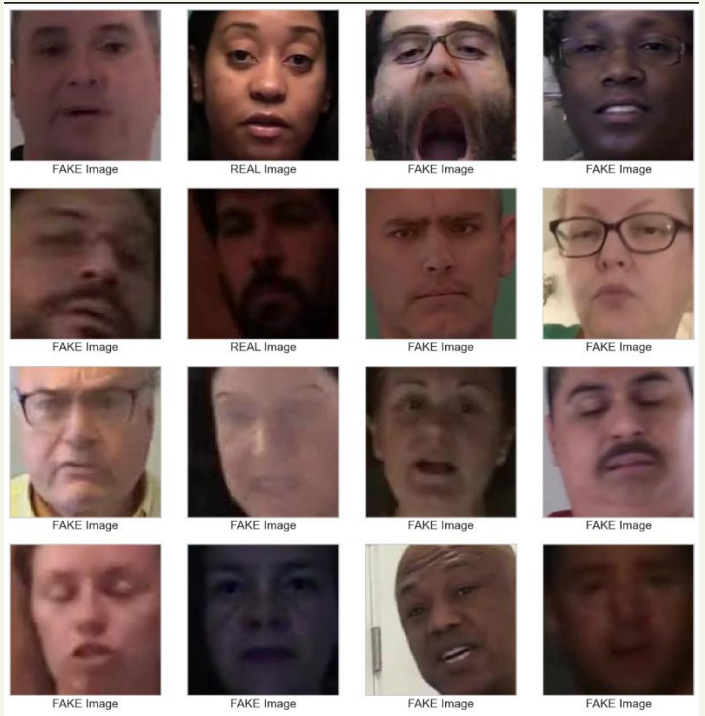
- More than 30,000 real images



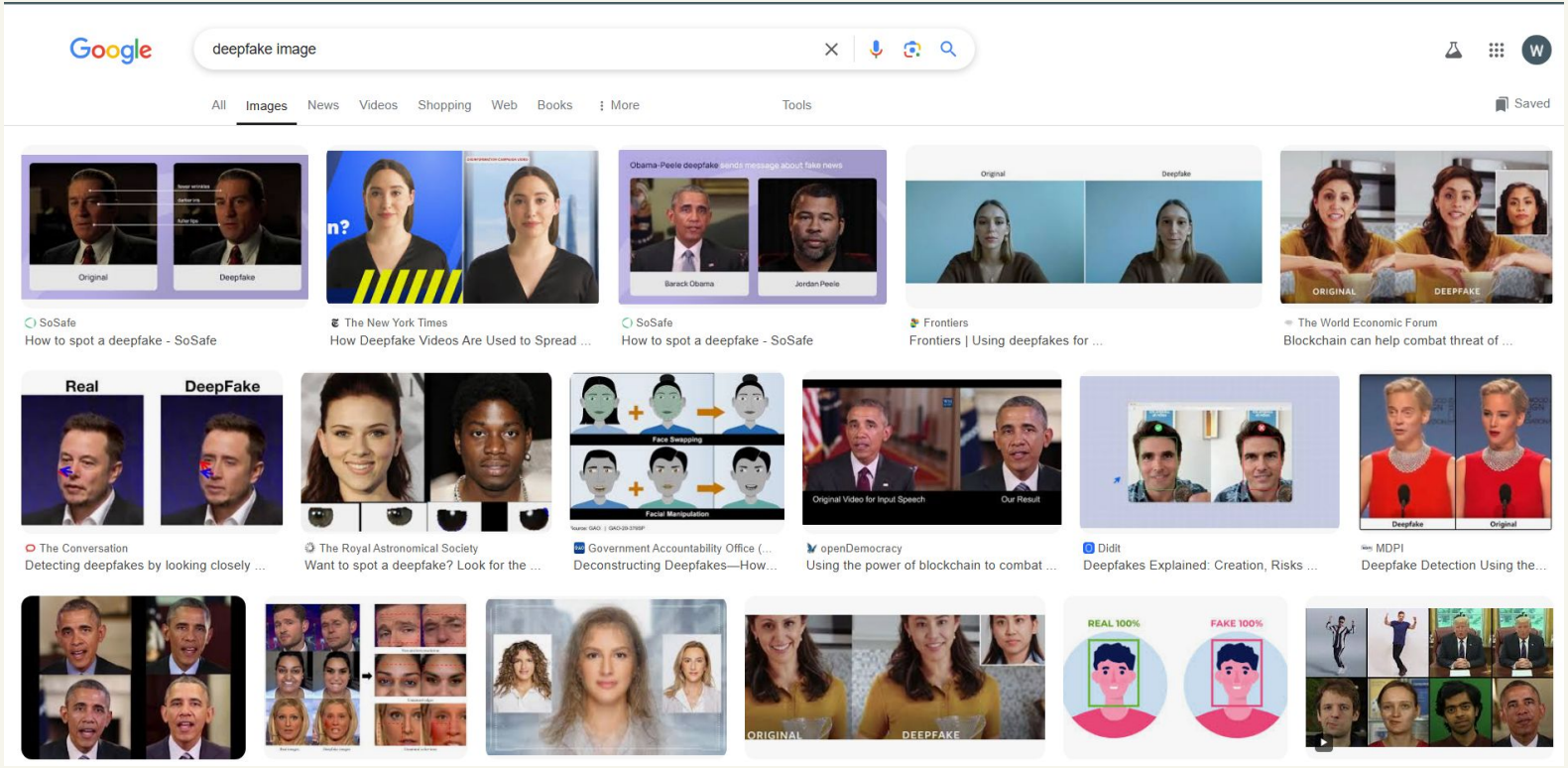
Data Example

Real:

Fake:



Scraping was not possible due to poor quality of images



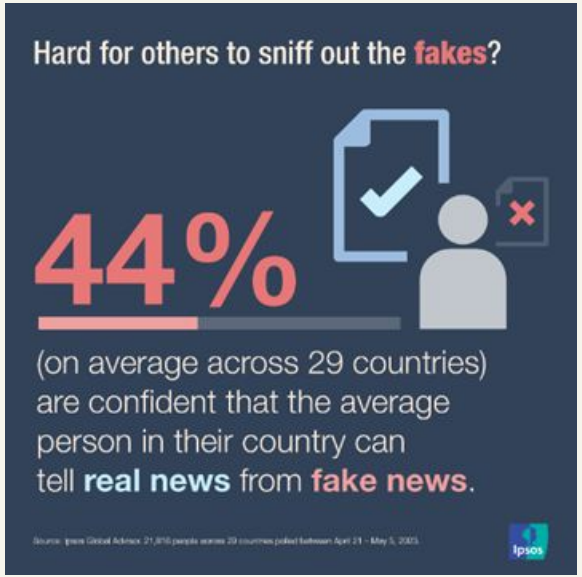
Identifying Fake News Online

By Jun Ming

Stats

The facts:

- 48 to 53 per cent - said they could tell if a piece of information on social media is true or false.
- However, about seven in 10 admitted that they have unknowingly shared fake news
- Not an issue limited to just singapore as across 29 countries less than half of the population are confident that they can identify fake news



THE STRAITS TIMES TECH

Many in Singapore confident they can spot fake news but may not actually be able to: Study

Example

- **Notable incident:**
  - Notable cases like Pizzagate conspiracy relating to Democratic party during 2016 campaign where democratic party was accused of child trafficking.
- Problem statement: How do we address online text misinformation with the rise of technology?

Pain Points

Panic among public

Unverified information creates a ripple effect of anxiety, leaving the public in a state of confusion and distress.

Impact Reputation

Damage the reputation of social media platforms and person/org/party the fake information is about.

Lack of trust

Misinformation undermines trust in institutions such as media, government, law enforcement or social groups. This could lead to divide in the country.

Solution

Key Issues:

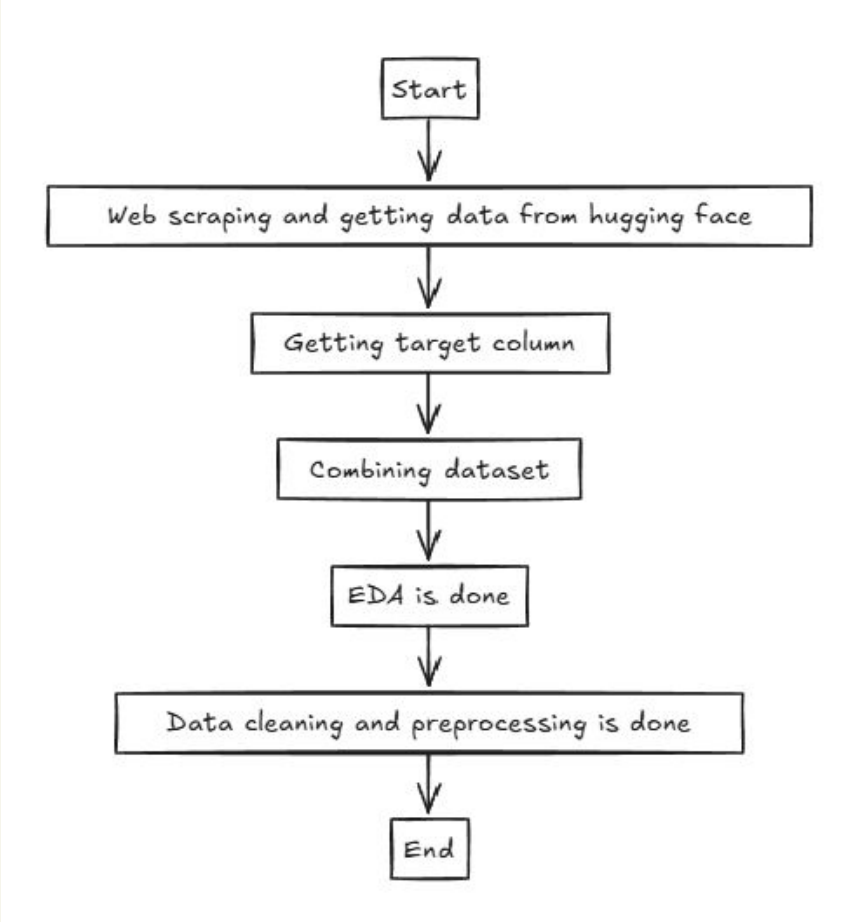
- Manual detection is **impractical, error-prone, and inconsistent**.

Fake news AI

- Using supervised learning technique, we can classify a statement made to whether it is fake news or not to increase the public's confidence in identifying fake news by 75%.

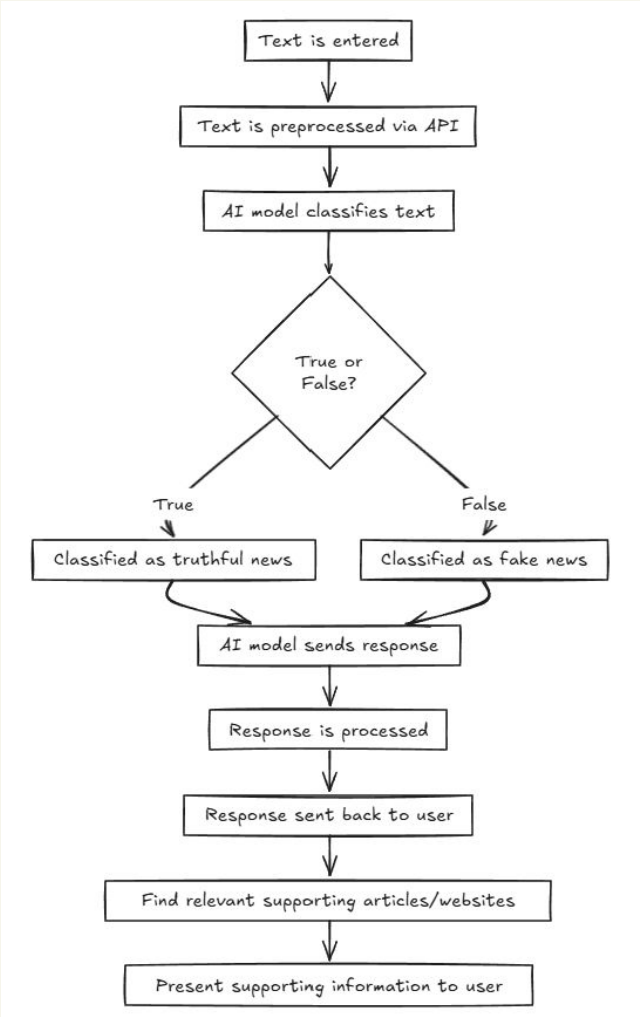
Target column:

- Half true and true are classified as true rest are False



Data Flow

- **Text Input:** User inputs text into the frontend.
- **Backend Processing:** Input is sent to the backend, where it undergoes consistent preprocessing steps.
- **Model Classification:** Preprocessed text is classified by the model.
- **Formatting:** The classification result is formatted.
- **Article Retrieval:** Relevant articles on the topic are retrieved.
- **Response:** The user receives the classified result along with the relevant articles.



Data

- **Web Scraping:**
  - Scraped **200 pages** from PolitiFact using Python.
  - Collected **~5K records**.
  - Includes 2023-2025
- **Hugging face Dataset:**
  - Retrieved the **LIAR2 dataset** from Hugging Face using Pandas.
  - Includes **~23K records**.
  - From 2000 to 2022
- **Final dataset:**
  - **~12K records**
  - From 2019 to 2025, and after cleaning.
  - 2 columns: statement and target column (0 for fake news 1 for truthful news)

	Unnamed: 0	statement_6	target
0	0	mitchell county north_carolina sheriffs deputi...	0
1	1	still havent crack thomas matthew crook phone	0
2	2	photo_shows child find several cities	0
3	3	state certify election	0
4	4	break second attack new orleans uncover police...	0
...	...	...	...
11893	11893	crimea sort take away president obama want rus...	0
11894	11894	suggest january traffic accident involve lab m...	0
11895	11895	image_shows ukrainian soldier pray	0
11896	11896	since minimum wage increase time raise time de...	0
11897	11897	quote rep kevin mccarthy say mass shoot japan ...	0

[11898 rows x 3 columns]

Objectives

Metrics

Develop a Fake News Detection Model with at least **75%** in F1 score which is a harmonic mean between precision and recall. Both fake news and true news detection are important

Data Manipulations

Use data visualization to identify issues (missing values, outliers, class imbalances) and perform data cleaning and preprocessing for model training.

Modelling

Test feature extraction methods (TF-IDF, CountVectorizer,Word2Vec, BERT), apply oversampling or class weighting, and evaluate multiple models and improve them. (Deep learning, ensemble learning, classical models)

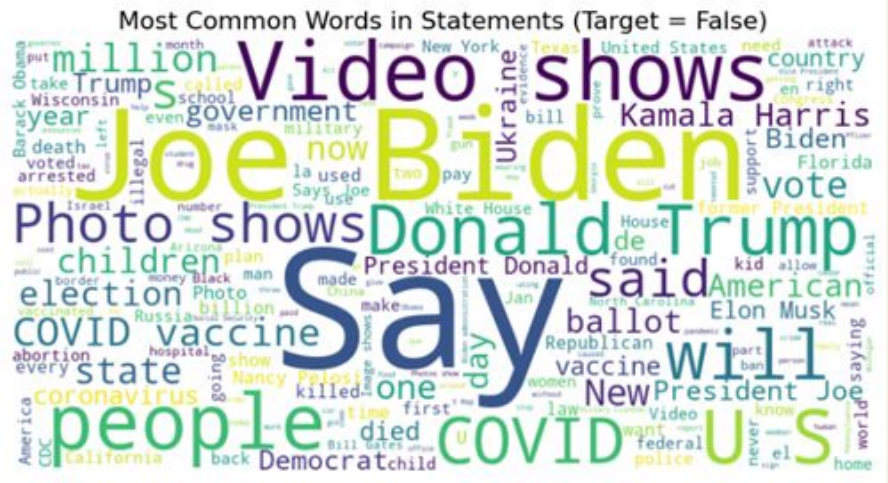
Integration

Integrate with other features and include article retrieval to help users make informed decisions.

EDA

Analyzing common words used in fake news

- The recent US election saw a spike in political fake news, along with common topics like COVID-19 and vaccination, indicated by politician names and terms like "COVID." Non-base word forms also appear, potentially increasing dimensionality.



Insights

Insights identified during EDA	Impact	Solution
Presence of stop words, punctuation, numbers, special characters, non-English sentences, and length outliers (1.5 IQR).	Increase dimensionality and introduce noise	Remove the records
Imbalance dataset	Skew the model towards majority class	Weighted class/Oversampling
Consistency within the texts	Increase dimensionality	• Lowercase everything • Lemmatize verbs • Filter out all columns except for statement and target • Remove not english records
Capturing context	Lack of context in the model's "understanding"	Get the top 100 bigrams and treat them as compound words and tokenize them together

Conclusion

- **Objective**: This project leverages AI to address key online safety challenges:
  - Hateful Memes
  - Mental Health Impacts of Harmful Content
  - Deepfake Misinformation
  - Fake News Detection
- **Impact**:
  - Provides data-driven insights
  - Enhances digital trust
- **Alignment with Singapore's Vision**:
  - Contributes to a secure, inclusive digital ecosystem
  - Empowers individuals and businesses
  - Fosters a sustainable digital future