IT3389 APPLIED AI PROJECT

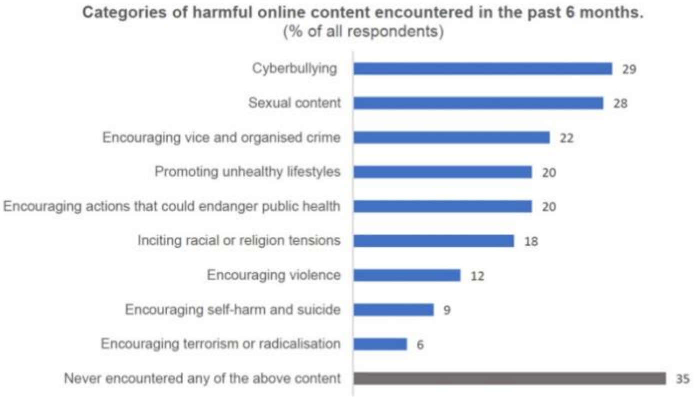
PROJECT PROPOSAL

Submitted By:	Gangula Karthik (223715Y)*
	Ng Jun Ming (220080D)
	Choy Wei Jun (221814H)
	Seah Pin Shien (220273H)
Team Name:	Team 3
Tutor:	Dr Veronica Lim
Submission:	14 January 2025

Introduction / Background Information

Online trust and safety is a critical global issue, and Singapore, as a highly digitalized nation, faces unique challenges in this domain. With a high internet penetration rate and a tech-savvy population, Singapore has implemented robust governance structures to address these challenges, including the Online Safety (Miscellaneous Amendments) Act and the establishment of the Centre for Advanced Technologies in Online Safety (CATOS). These initiatives have strengthened the nation's capacity to detect and counter harmful content, reinforcing its commitment to creating a secure and trustworthy digital ecosystem. However, the rapidly evolving landscape of technology, including AI-generated content like deepfakes, and the diverse nature of online threats continue to complicate enforcement efforts.

Despite Singapore's progress, statistics reveal significant concerns about online safety. Nearly 60% of children in Singapore experienced at least one form of online harm in 2023, slightly below the global average of 67%, yet still troubling. Similarly, two-thirds of internet users reported encountering harmful content, with cyberbullying and sexual content being the most prevalent. Alarmingly, many users take no action when exposed to such content, with only 25% reporting incidents to platforms, often facing difficulties in the process [15].



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

Figure 1 – Categories of harmful online content

As an AI startup addressing sustainability issues, we emphasize social sustainability, which promotes inclusive, equitable, and safe environments to enhance societal well-being and trust. In the digital realm, this means fostering online spaces where individuals feel respected, protected, and empowered to engage without fear, aligning with UN sustainability development goals 11 (sustainable cities and communities) and 16 (peace, justice and strong institutions). By prioritizing online trust and safety, this project supports Singapore's vision of a digitally inclusive society. Leveraging advanced technologies and data-driven solutions, it complements initiatives like CATOS, bridges enforcement gaps, and strengthens digital cohesion, ensuring citizens can confidently participate in the digital economy.

This project aims to tackle these challenges by analyzing online content and behavior data to develop predictive models for identifying trends and potential threats. By delivering actionable insights, the initiative will empower **social media platforms, content creators, and businesses** to implement effective countermeasures and create safer online spaces. These efforts align with Singapore's vision of building trust in the digital economy while enhancing user engagement and confidence, ultimately benefiting businesses, society, and the broader digital community.

Business Scenario / Proposed Solution

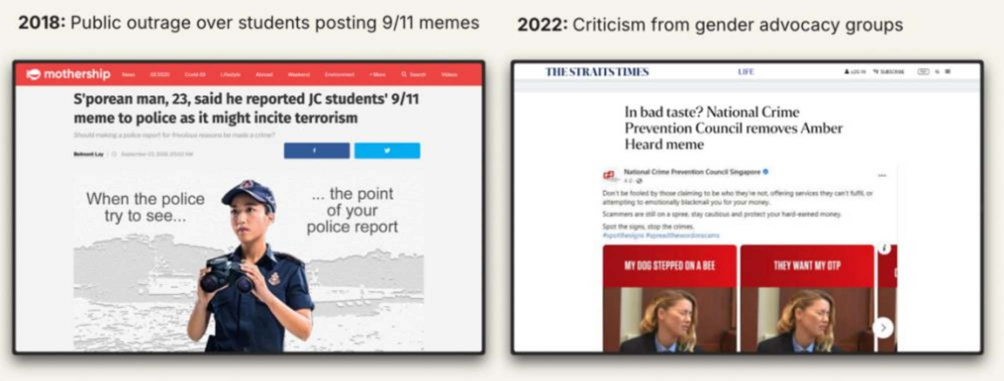
In today's digital age, the prevalence of harmful online content poses significant challenges to creating safe and trustworthy digital environments. While many individuals and organizations recognize the importance of fostering online trust and safety, tangible actions to address these challenges often lag. This gap is particularly pressing in a highly connected society like Singapore, where the consequences of unchecked digital threats can have far-reaching social and economic implications. Our team is strategically focused on addressing four critical areas: **Hateful Meme Classification**, **Social Media Post Sentiment Analysis**, **Deepfake Identification**, and **Fake News Detection**. Each area contributes to building a safer, more inclusive, and sustainable digital ecosystem. Empathy and user-centered solutions are crucial to understanding and addressing these digital threats effectively. Several concerns have been identified:

Classification of Hateful Memes Online (Karthik):

Hateful meme classification involves identifying and mitigating harmful content embedded in memes that target racial, religious, and political sentiments. Memes, a dominant form of online communication, combine visual and textual elements to convey humor or satire. However, they are increasingly exploited to disseminate hateful content, posing significant societal harm. Addressing hateful memes is critical to fostering inclusive and respectful online environments. The proliferation of hateful memes can have serious consequences, including the normalization of hate speech, reinforcement of stereotypes, and potential incitement to real-world violence. As these memes make their way from fringe communities to mainstream platforms, they can contribute to the radicalization of individuals and the toxification of public discourse [17].

Pain Points:

- **Societal Harm and Divisive Content:** Approximately 60% of memes contain divisive or harmful messages [1]. Such content perpetuates hate, misinformation, and violence, intentionally or unintentionally. For example, an Instagram post by Nanyang Junior College students in 2018 trivialized the 9/11 attacks, sparking public outrage, a shutdown of the page, and an apology [3]. Similarly, in 2022, the National Crime Prevention Council (NCPC) faced backlash for posting an Amber Heard meme during her trial, prompting criticism from gender advocacy groups and its eventual removal [4].



2018: Public outrage over students posting 9/11 memes

S'porean man, 23, said he reported JC students' 9/11 meme to police as it might incite terrorism

Should making a police report for frivolous reasons be made a crime?

mothership
When the police try to see...
... the point of your police report

2022: Criticism from gender advocacy groups

In bad taste? National Crime Prevention Council removes Amber Heard meme

National Crime Prevention Council Singapore

Don't be fooled by those claiming to be who they're not, offering services they can't fulfil, or attempting to emotionally blackmail you for your money.

Scammers are still on a spree, stay cautious and protect your hard-earned money.

Spot the signs, stop the crimes.
#spotthesigns #spreadthewordonscams

Figure 2 – Examples of memes leading to outrage in Singapore

- **Complexity of Multimodal Content:** The combination of visual and textual elements in memes complicates detection efforts, making it challenging to discern harmful intent or hidden messages embedded in the content.

By leveraging advanced machine learning techniques for multimodal analysis, we can detect hateful memes effectively and address their societal impact. This fosters safer online environments, reduces exposure to harmful content, and ensures a more inclusive digital space.

Predicting Impact of Social Media Text on Mental Health (Pin Shien):

Social media has become a pervasive part of daily life, offering numerous benefits such as connectivity and information sharing [5]. However, it has also introduced significant challenges, including the rise of harmful content that adversely affects mental health [5]. A government survey in Singapore revealed that 74% of respondents encountered harmful content online, a sharp increase from 65% in the previous year [6]. This harmful content includes cyberbullying, sexual content, and a worrying rise in posts inciting racial or religious tensions, as well as violent material [6].

Despite these alarming trends, action against such content remains limited. Only 25% of respondents reported harmful content to platforms, while 33% blocked offending users [6]. Alarmingly, 60% of individuals chose to ignore such content, leaving its potential negative impact unaddressed. [6]

The prevalence of harmful content, coupled with its underreporting, poses a growing risk to mental health, exacerbating issues like anxiety, depression, and emotional distress. The lack of actionable insights into how exposure to such content correlates with mental health outcomes further complicates efforts to mitigate these effects.

Pain Points:

- **Algorithmic Amplification of Harmful Content**: 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 susceptible to the mental health effects of harmful content, increasing risks of long-term psychological issues.

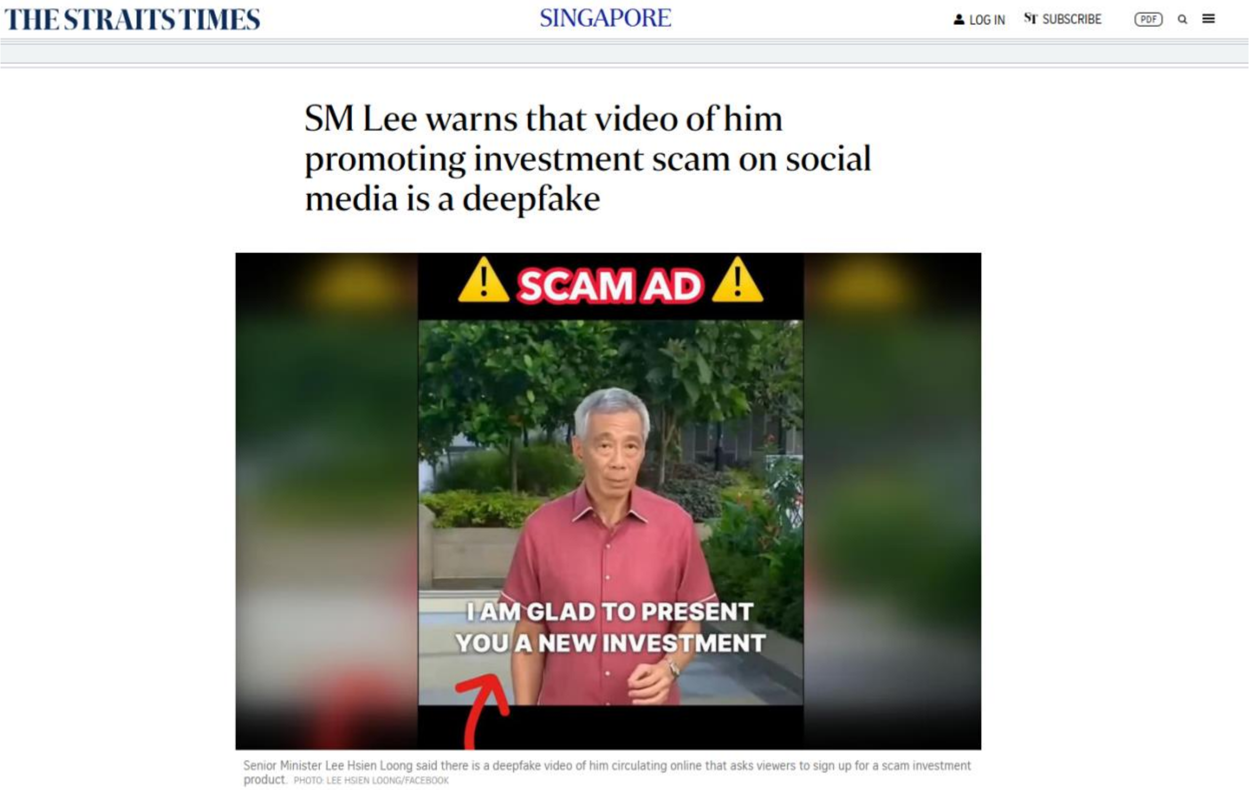
This sub-topic aims to analyze and predict the impact of social media usage on mental health by examining patterns of harmful content exposure, user responses, and their psychological consequences. Leveraging data analytics and machine learning, the goal is to identify risk factors, understand behavioral responses, and develop predictive models to inform effective interventions, ensuring healthier and safer social media experiences for individuals.

Detection of Deepfake Misinformation (Wei Jun):

Addressing the Rise of Deepfake Misinformation with AI-Powered Detection

Deepfake technology poses a growing threat due to its ability to create highly convincing fake videos, which are often used to spread false information, damage reputations, and facilitate scams. With deepfake fraud surging by 1,740% in North America and 1,530% in Asia-Pacific in 2022, the rapid rise of this technology shows its increasing exploitation for malicious purposes. Additionally, more than 95% of deepfake videos are created using DeepFaceLab [9], an open-source software that is easily accessible and can be misused. One of the biggest concerns is the inability of many people to confidently identify cloned voices, with 70% of individuals unable to tell the difference between a real voice and a manipulated one, making them vulnerable to scams like voice phishing.

A prominent case highlighting the dangers of deepfakes involved the use of a manipulated video featuring Singapore’s Prime Minister, Lee Hsien Loong, promoting an investment scam on social media. This deepfake video was so realistic that it deceived many viewers, contributing to a loss of trust and raising concerns about the potential for future manipulation of public figures for malicious purposes



THE STRAITS TIMES SINGAPORE

SM Lee warns that video of him promoting investment scam on social media is a deepfake

Senior Minister Lee Hsien Loong said there is a deepfake video of him circulating online that asks viewers to sign up for a scam investment product. PHOTO: LEE HSIEN LOONG/FACEBOOK

Pain points:

- **Erosion of Trust**: Deepfakes undermine people's trust in the media and public figures, as it becomes increasingly difficult to distinguish between real and fake content.
- **Increased Vulnerability to Scams**: With the rise of deepfake technology, people are more susceptible to scams, such as voice phishing and identity theft, as fake voices and videos become harder to detect.
- **Public Safety Concerns**: Deepfakes can be used to manipulate public opinion or incite violence, creating social and political instability by spreading false information about individuals, governments, or events.
- **Legal and Ethical Challenges**: The creation and distribution of deepfakes raise serious legal and ethical questions, as victims may struggle to find recourse against malicious actors who use these videos for defamation, fraud, or harassment.

Identification of Fake News Online (Jun Ming):

In recent years, the proliferation of fake news online has become a significant concern. Notable examples include the "Pizza gate" conspiracy theory in 2016, which falsely alleged that Hillary Clinton and members of the Democratic Party were involved in a child trafficking ring operating out of a Washington, D.C., pizzeria.[13] This baseless claim gained traction on social media, fuelled by misinformation and viral posts, despite being thoroughly debunked.

The controversy had real-world consequences, including the spread of distrust in political institutions and the endangerment of lives. For instance, a man entered the pizzeria armed with a rifle, claiming he was there to "self-investigate" the allegations. The incident not only exposed the dangerous ripple effects of fake news but also cast a shadow over Clinton's 2016 presidential campaign, contributing to the polarization of voters and undermining public discourse. [13]

In Singapore, 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 - 69 to 76 percent - admitted that they have unknowingly shared fake news. This is not limited to Singapore, a survey conducted across 25 countries revealed that 86% of online global citizens believe they have encountered fake news at some point, with nearly 90% initially accepting it as true before later discovering otherwise.[14].

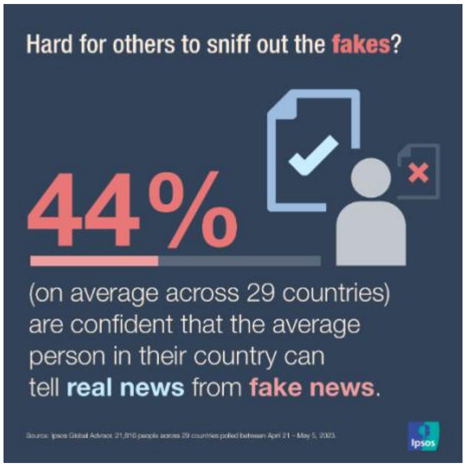

Hard for others to sniff out the fakes?
44%
(on average across 29 countries) are confident that the average person in their country can tell real news from fake news.
Source: Ipsos Global Advisor. 21,816 people across 29 countries polled between April 21 - May 5, 2023.
Ipsos

Figure 3 – Summary of survey results

Moreover, many people lack confidence in distinguishing fake news, with some overestimating their ability to identify it. To address this issue at its core, a robust, consistent, and automated solution is needed to efficiently identify and mitigate the impact of fake news in real time.

The increasing volume of online discussions makes it impractical to manually detect and moderate fake news in comment sections.

- **Manual efforts are often prone to human error and inconsistencies**, leading to delays in addressing false information.
- **This unchecked spread of fake news damages reputations**, causes unnecessary panic, and undermines trust in digital platforms.
- Many people **lack confidence in distinguishing fake news**, with some **overestimating their ability** to identify it.

Scope of Work

In this section each of the members have documented the modelling objectives as well as the exploratory data analysis and data preparation steps in detail.

Classification of Harmful Memes Online (Karthik)

Solution

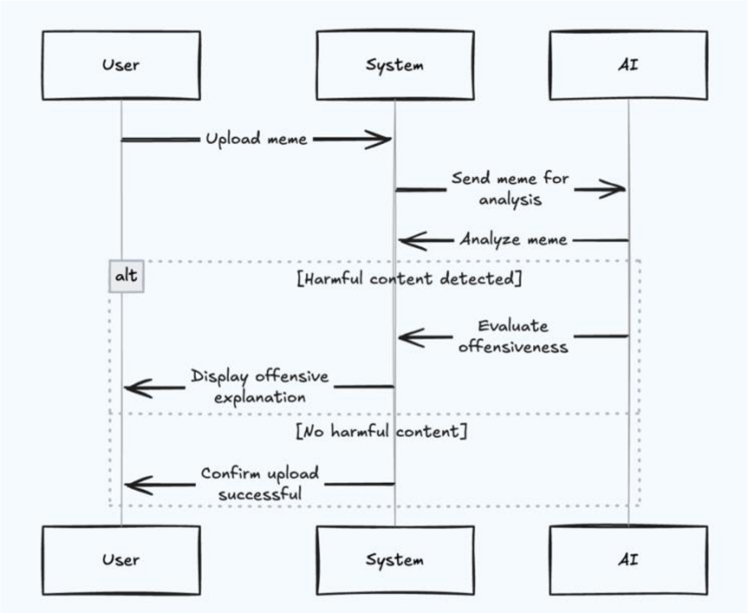


Figure 4 – Diagram of hateful meme detection system

- A multimodal classifier that detects if a meme is harmful or not.
- If it is harmful, explain to the user why it is.
- Otherwise, it can be uploaded onto the social media platform

Objectives of Project

The primary objective is to develop a multimodal classifier on memes, utilizing both images as well as the textual captions provided on the image. The classifier will then be able to classify whether it is a harmful meme or not. To achieve these the below need to be performed:

1. **Data Visualizations and Insights:** Both the textual and image data from the memes will be visualized and cleaned to identify any errors.

2. **Experimentation with Models:** Various embedding models and different neural network architectures will be considered alongside classical machine learning models as well to see which one is able to generalize to the problem statement better. Hyperparameters will also be tuned to achieve optimal model performance.

3. **Performance benchmarks:** The model will use AUROC, a metric well-suited for datasets often affected by class imbalance and tells how well the model is able to separate the 2 classes. A higher AUROC indicates stronger discriminative performance, making it ideal for comparing different models and approaches. Widely used in other hateful meme classification papers, AUROC is also an official evaluation metric in challenges like the Facebook Hateful Meme Challenge, reinforcing its relevance and reliability.

4. **Additional Feature:** Due to the opaque, black-box nature of the complex classifiers, there is need to use explainability techniques. Vision language models such as Pali gemma, Llava-instruct or smol-VLM can be prompted to explain what exactly is offensive in the meme and give suggestions on how to make it less offensive.

Dataset Used

The dataset comes from the Facebook harmful memes challenge. It contains 2 different classes which are labelled as harmful and not harmful. The dataset was severely imbalanced as well, so web scraping was used to gather data from the internet (Bing search) to balance the harmful class. 100+ offensive topics were fed into the data scraping tool to extract approximately 2000 images in total.

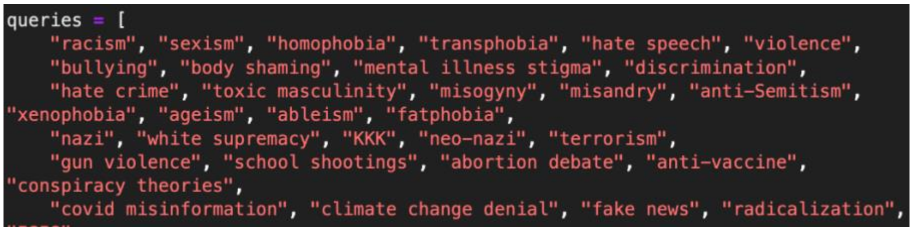
Link to the dataset: https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset/data

Link to the data scraping tool: https://github.com/ostrolucky/Bulk-Bing-Image-downloader

Web Scraping Images

The images were scrapped using the Bing search image downloader API which can be downloaded through the python package manager. However, the images will still need to be cleaned, and the meme text will need to be extracted before it can be merged with the main dataset, which is the Facebook harmful memes dataset. This step will cover the steps and considerations taken to get the raw merged dataset.

The first step is to gather the images based on the different offensive topics. The 100+ offensive topics were generated using ChatGPT and cross checked by the developer to ensure that valid topics were chosen. Those which were not valid were removed from the list.



```
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", "ableism", "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",
```

Figure 5 – queries of hateful meme topics

After that the top 20 image results from each topic were chosen and in total 2,668 images were downloaded through web scraping.

The next step that was done is to remove the abnormally large images, from the images that were scraped, some of them contained posters as well which were not really memes. These were very long in height and width. When resizing extreme outlier images to fit within a standard size, the aspect ratio or key features might be distorted and cropping them will lead to a loss in information. Instead it is easier to remove them from the dataset.

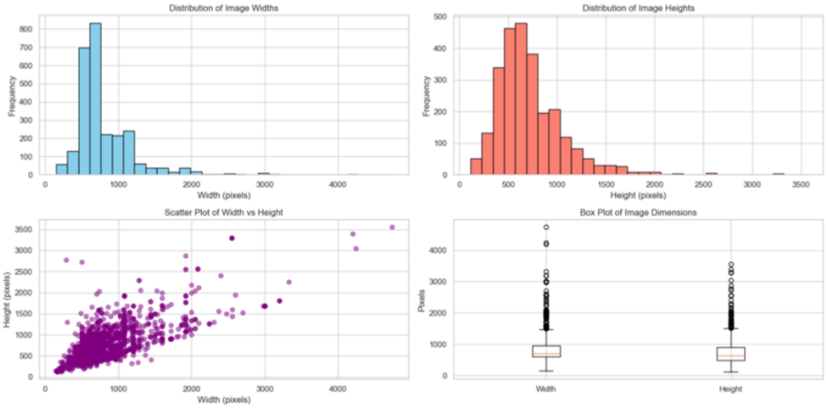


Figure 6 – Identifying outlier images

Those images with a width more than 2000 and a height more than 2500 were removed from the dataset. After this, the duplicate images were also removed from the dataset since it can bias the model, giving more weight to that image and lead to overfitting. The duplicate images were found through hashing the images using md5, if the hash of the images matched, it meant that the images were the same and they needed to be removed. Below are some examples of the duplicated memes:

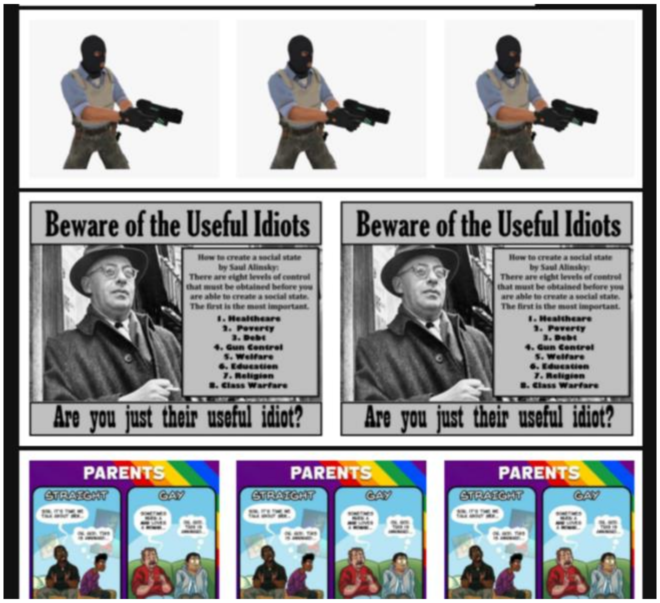

Beware of the Useful Idiots
How to create a social state
by Saul Alinsky:
There are eight levels of control
that must be obtained before you
are able to create a social state.
The first is the most important.
1. Healthcare
2. Poverty
3. Debt
4. Gun Control
5. Welfare
6. Education
7. Religion
8. Class Warfare
Are you just their useful idiot?
Beware of the Useful Idiots
How to create a social state
by Saul Alinsky:
There are eight levels of control
that must be obtained before you
are able to create a social state.
The first is the most important.
1. Healthcare
2. Poverty
3. Debt
4. Gun Control
5. Welfare
6. Education
7. Religion
8. Class Warfare
Are you just their useful idiot?
PARENTS
STRAIGHT
GAY
PARENTS
STRAIGHT
GAY
PARENTS
STRAIGHT
GAY

Figure 7 – Removing duplicate images

The next step is to extract the text which is embedded on the meme image. This is a challenging task since traditional OCRs struggle at recognizing the text when there is a noisy background. Additionally, there are watermarks on some of the scraped memes which means these will need to be removed during preprocessing so that the OCR does not pick these up.

For this image processing techniques such as adjusting the contrast and image thresholding were used on the image to create a mask and image morphology was used to identify and remove small objects, such as these watermarks.

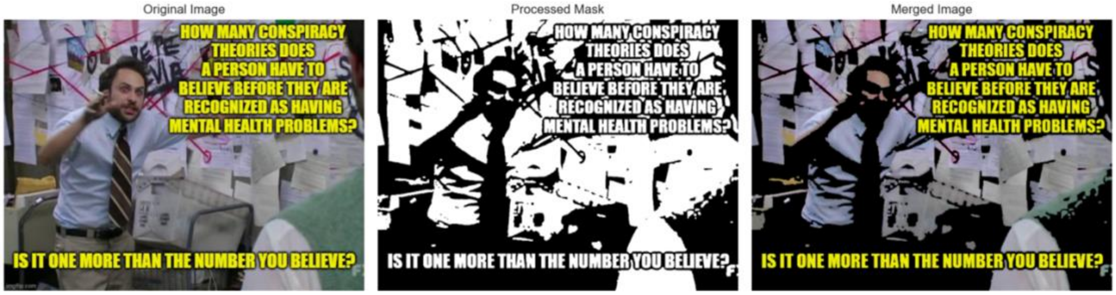


Figure 8 – Ineffective image processing

The above example has a small watermark on the bottom left but it has been removed successfully. For the next step a bilateral filter and color adjustments were applied to dim the background and highlight the text.

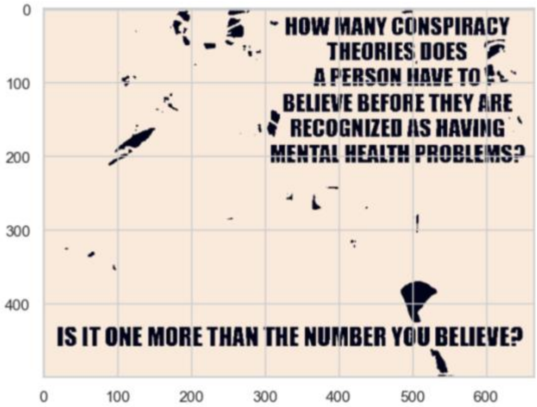
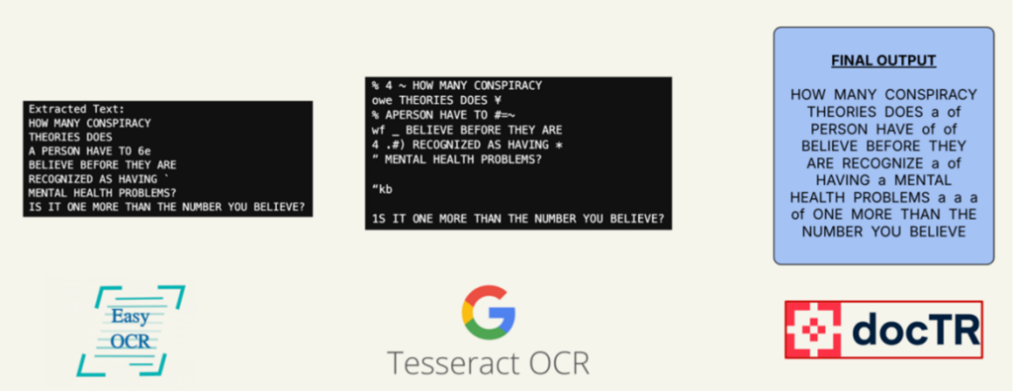


Figure 9 – residual mask left behind on images

Once the text was able to stand out against the rest of the image, as in the above example, the OCR could then be applied. 3 different OCRs were experimented with all 3 showing poor/inaccurate results. This was due to the image processing techniques unable to generalize on all images, leading to poor results. Below are the results of all the OCR models used:



```
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

Figure 10 – Experimenting with different OCR techniques

To get consistent accurate results, with minimal or no human annotation was not possible with these OCR models, especially given the lack of time. Hence, vision language models (VLM) were used for this task. These models are known to be far more accurate than traditional OCR methods used above. For this use-case the Smol-VLM is used given its low RAM usage and small size compared to other models. The model can also be run on a consumer GPU such as the T4 GPU available on the google colab free tier.

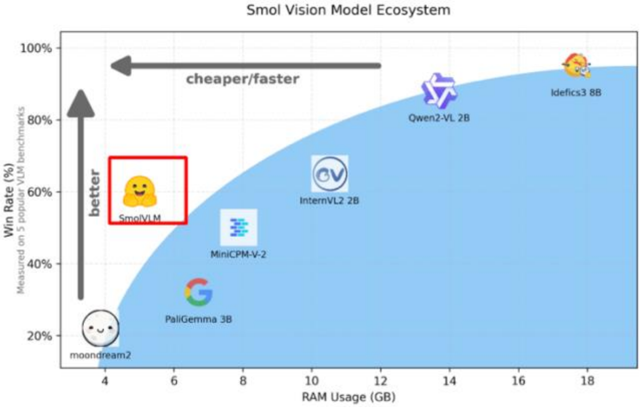
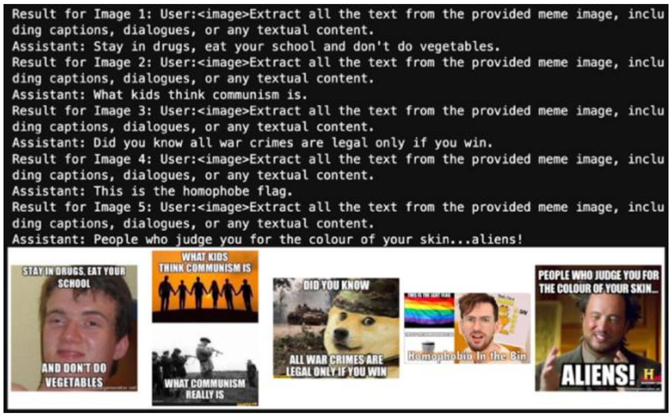


Figure 11 – Results of Smol-VLM performance

After finding a suitable prompt for the VLM OCR task, it was tested on multiple images. Where it was found to have extremely high quality results as compared to the previously used OCR models.

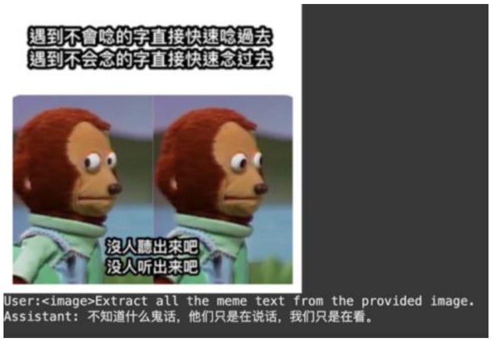


```
Result for Image 1: User:<image>Extract all the text from the provided meme image, inclu
ding captions, dialogues, or any textual content.
Assistant: Stay in drugs, eat your school and don't do vegetables.
Result for Image 2: User:<image>Extract all the text from the provided meme image, inclu
ding 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!
```

STAY IN DRUGS, EAT YOUR SCHOOL
AND DON'T DO VEGETABLES
WHAT KIDS THINK COMMUNISM IS
WHAT COMMUNISM REALLY IS
DID YOU KNOW
ALL WAR CRIMES ARE LEGAL ONLY IF YOU WIN
Homophobia In the Bin
PEOPLE WHO JUDGE YOU FOR THE COLOUR OF YOUR SKIN...
ALIENS!

Figure 12 – Zero shot OCR using Smol-VLM on meme dataset

The VLM was also tested on memes which were in other languages, where it also produced similar consistent results, based on the feedback from my Chinese speaking classmates.



```
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。
```

Figure 13 – Results of SmolVLM on multi-lingual texts

All the results were written to the csv file and the code for the VLM-OCR was written in smol-VLM_ocr_extraction.ipynb file since it was run on google colab.

Some code implementation details to highlight would be:

- With code optimizations for low GPU memory utilization and garbage collection on every iteration, ensured that the google colab memory limit was not hit which would cause the kernel to crash.
- The progress was also saved in the csv file so that if execution halted for whatever reason, it would continue from the last processed record in the dataset.

Exploratory Data Analysis

Summary of General Dataset Understanding:

Title	Description
Distribution of the target column	• Target column is binary (Hate or Not Hate). • It shows data imbalance in a ratio of 64% to 36%. • After merging the dataset with web scraped images, the data is balanced.
Duplicate and null values check in Facebook harmful meme dataset	• Checked to see if there are any null / duplicate values in the dataset. • No duplicate or null values were identified in the dataset.
Distribution of the File Formats	• Most of the images were PNG and JPG • There were some GIFs (animated images) and only one BMP (large, raw, high-quality images) files, these were removed from the dataset.

Summary of Textual Data Understanding:

Title	Description
Language Detection *(using Lingua 2.0 [high accuracy mode] python package)*	• Majority of the texts were in English • The best open source language detector was used, refer to appendix [image 1]. • A small portion of the images (around ~4%), were in another language. While the detection was not 100% accurate due to

	presence of non-English words in memes, these were removed to ensure there is no confusion in the modelling stage.
Sentence length vs average word length	• A scatterplot of the average sentence length and average word length per sentence was plotted to see how long the sentences and words are. • Those records with more than 100 words per record were removed, as usually the memes are very short and don't contain a lot of text. • The words with more than 50 characters were also removed as these would be considered as outliers.
Top punctuations count by target label	• A group bar chart was plotted to identify which punctuation type is more common among the target labels. • Hate speech often includes symbols like "$" and "%" (suggesting financial or numerical references) and quotation marks, possibly used for humor or satire of notable individuals. • Non-hate speech surprisingly contains more exclamation marks ("!"), typically linked to strong emotions, defying the expectation that such marks are more common in hate speech.
Distribution of usernames, emails and hashtags	• Like the above analysis, a group bar chart was plotted. • Hate memes include more numbers, usernames, and hashtags • Usernames will be redacted from the training data to maintain privacy.
Identification of Websites	• Memes with large website watermarks were manually inspected; images with only website names as text were removed. • Contextual website references (e.g., "ancestry.com" vs. "match.com") were retained as they are crucial for identifying harmful memes.
Parts of Speech Tagging by target label	• A group bar chart was used here as well • Hateful memes use nouns, adjectives, and adpositions more frequently, emphasizing specific entities, detailed descriptions, and explicit relationships. • This increased use of structured and descriptive language suggests a focus on people, places, or things, often creating stronger connections between them.
Named Entity Recognition by target label	• A heatmap was used here to plot the different entities by the target label. • Hate memes frequently reference specific entities like PERSON, NORP, ORG, and GPE, reflecting their focus on real-world people, groups, and locations, often with political or social implications.

	• The mention of MONEY and PERCENT in hate memes suggests the use of exaggerated statistics or sensationalized narratives, contrasting with the more general and factual nature of non-hate memes.
Frequent and less frequent words	• Hate memes contain significantly more race-related terms (e.g., "Asian," "Muslim," "white") than non-hate memes, despite both categories including swear words.

Summary of Image Data Understanding

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.
Color Scale and Dominant Color Analysis by target label	• Color scales in hate and non-hate meme images are similar, with minimal impact on prediction. • Looking at the dominant colors, non-hate memes tend to have slightly darker shades than hate memes.

Data Preparation

In this section, the aim is to clean and prepare the data in a way that is easily understood by the model for modelling.

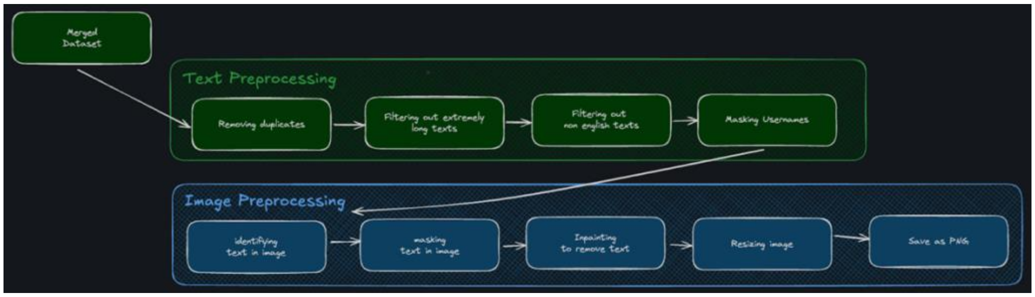
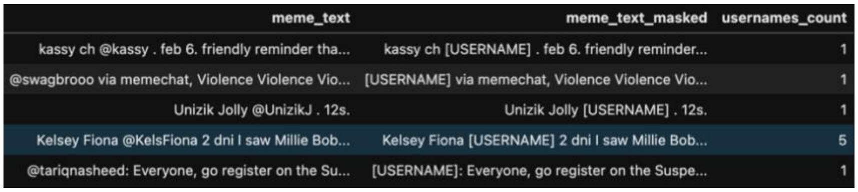


Figure 14 – Data preparation pipeline

Below are the steps that will be taken in the data preparation phase:

1. **Ensuring consistent image formats of PNG**
2. **Removing the duplicate images by comparing image hashes**
3. **Removing of excessively long meme texts (greater than 100 words)**
4. **Removal of memes with excessively long words (greater than 50 characters)**
5. **Masking of the usernames to ensure privacy of individuals**; Names however were not masked since it could be used to identify the hate. For example – “Hitler” and “Osama Bin Laden” are

names which are usually linked to a negative sentiment. Below is an example of the masking process:



meme_text	meme_text_masked	usernames_count
kassy ch @kassy . feb 6. friendly reminder tha...	kassy ch [USERNAME] . feb 6. friendly reminder...	1
@swagbrooo via memechat, Violence Violence Vio...	[USERNAME] via memechat, Violence Violence Vio...	1
Unizik Jolly @UnizikJ . 12s.	Unizik Jolly [USERNAME] . 12s.	1
Kelsey Fiona @KelsFiona 2 dni I saw Millie Bob...	Kelsey Fiona [USERNAME] 2 dni I saw Millie Bob...	5
@tariqnasheed: Everyone, go register on the Su...	[USERNAME]: Everyone, go register on the Suspe...	1

Figure 15 – masking of the username

6. **Removal of texts which contain other languages**
7. **Masking text on image and doing inpainting:**

For the final preprocessing on the images, the meme text was removed from the image. To do this, first the meme text region of interest was identified using the easyOCR module. Once the text was identified the mask was generated and the image inpainting was done.

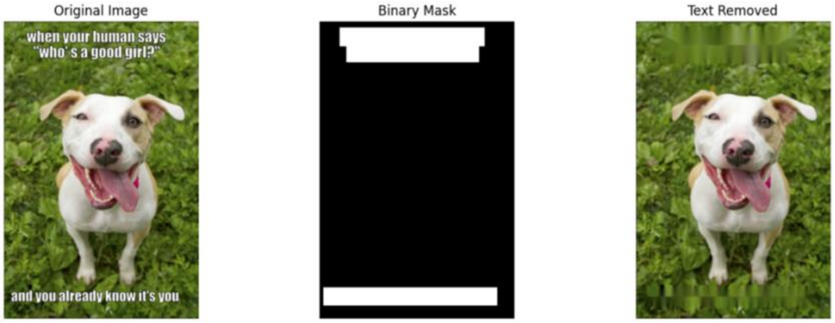

Original Image
when your human says "who's a good girl?"
and you already know it's you
Binary Mask
Text Removed

Figure 16 – Data preprocessing on the images

Finally, all the images were resized to a size of 512 x 512 and saved. The aspect ratio was maintained by center cropping and padding the images.

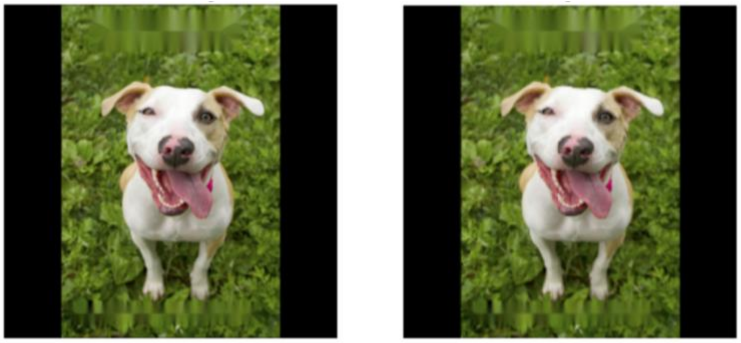


Figure 17 – Cropping of the images

Classification of Fake News Online (Jun Ming)

Solution

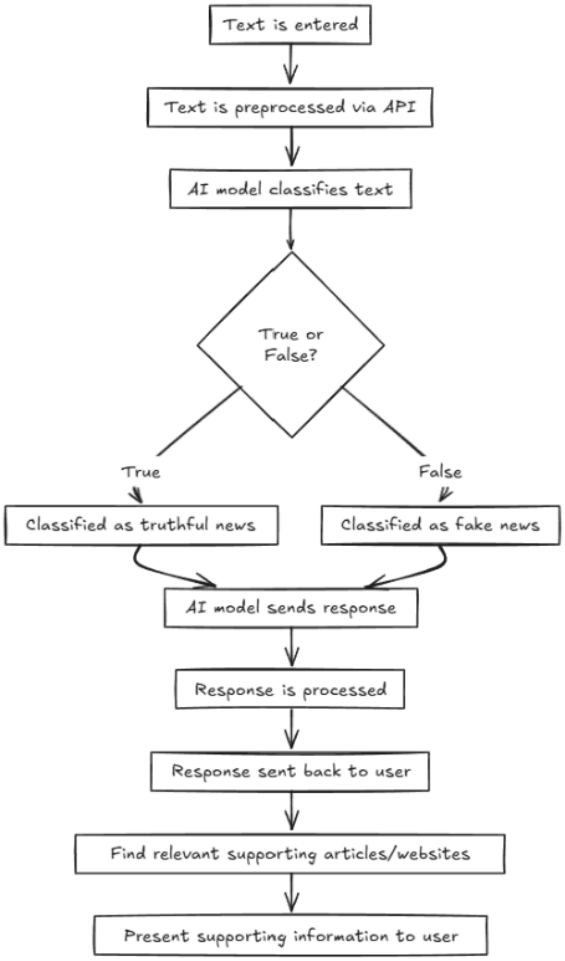

Text is entered
Text is preprocessed via API
AI model classifies text
True or False?
True
False
Classified as truthful news
Classified as fake news
AI model sends response
Response is processed
Response sent back to user
Find relevant supporting articles/websites
Present supporting information to user

Figure 18.1 – User flow for fake news detection

The primary objective is to develop a supervised learning model that classifies statements as either fake news or truthful.

Workflow is as follows:

1. The input text is submitted through the frontend and sent to the backend for processing.
2. The backend applies the same preprocessing steps to the text as used during model training.
3. The model classifies the text and generates a response.
4. The system formats the response and retrieves relevant articles on the topic.
5. Finally, the classified result, along with the retrieved articles, is returned to the user.

This provides the user with transparency as they can verify the classification results against the retrieved articles, fostering trust and offering context for the decision made by the model.

Objectives of Project

Objectives:

The main goal is to develop a Fake News Detection Model. The objective is to design and implement an AI-based model capable of analyzing and classifying statements as genuine or fake with at least 80% accuracy (or other key metrics such as recall or precision). This tool will help users make informed decisions about the veracity of statements.

1. **Data Preparation and Visualization**
   - Leverage data visualization techniques to identify insights and potential issues within the dataset (such as missing values, outliers, and class imbalances).

- Perform necessary data cleaning and preprocessing, ensuring the dataset is ready for model training.

2. **Model Experimentation and Evaluation**
    - Experiment with various feature extraction techniques such as TF-IDF and Count Vectorizer for text preprocessing.
    - Evaluate different supervised learning models, starting with simpler models like Logistic Regression and exploring more complex models like Random Forests and ensemble methods to determine the best-performing model.
3. **Enhancing Model Support and Transparency**
    - Integrate a feature that fetches external sources from the internet, providing additional context to the classification results. This feature will allow users to verify the model's output and help them make more informed decisions by cross-referencing the prediction with credible sources.

Dataset Used

The flow chart below describe how the data is derived as well as the steps taken to reach the end data which is cleaned and preprocessed.

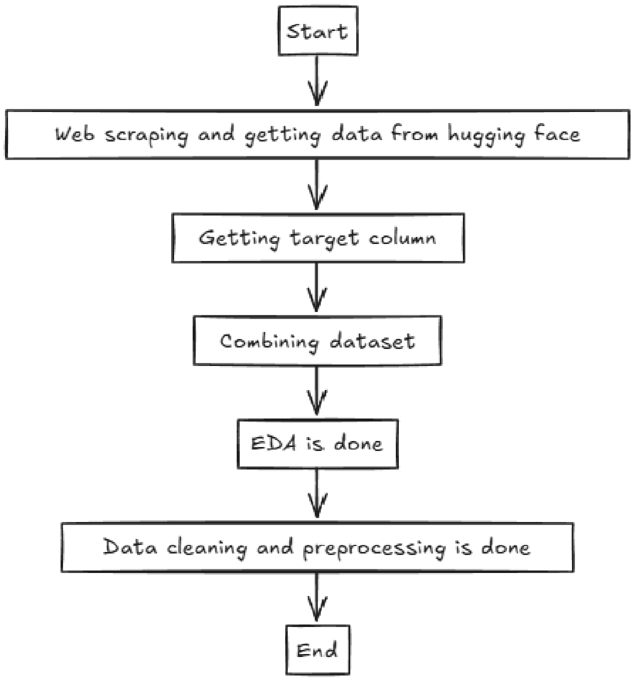
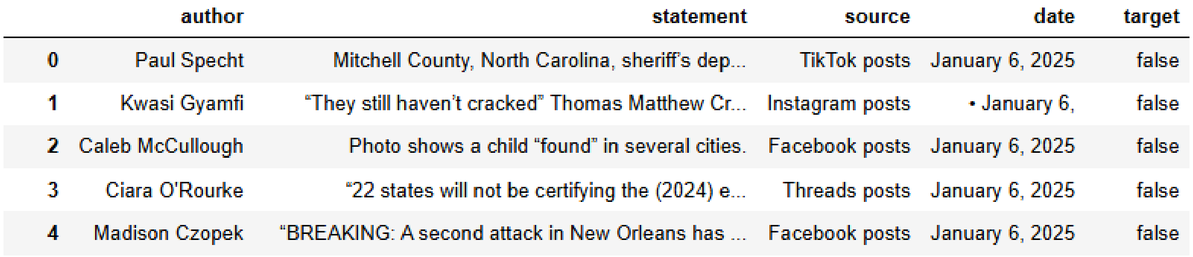

Start
Web scraping and getting data from hugging face
Getting target column
Combining dataset
EDA is done
Data cleaning and preprocessing is done
End

Figure 18.2 – Data Processing pipeline for fake news

The dataset used for this project is from web scraping and from an existing dataset, which is publicly available at Liar2 Dataset and the website scraped is Politifact. The code to derive these data is submitted individually. Both datasets, PolitiFact and Liar2, are in English, so there is no need to check for other languages.

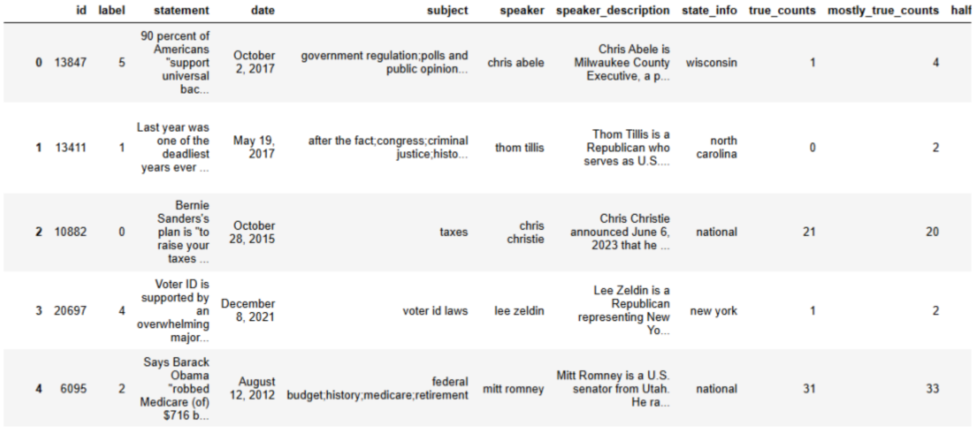
- Polifact contains news content with labels annotated by professional journalists and experts, along with social context information. Web scraping was used to gather data, scraping 200 of the latest pages resulted in around 5K records. The data period is from 2023 to 2025. Contains the columns: author, 'date', 'statement', 'source', 'target'. Image below shows the example data.



	author	statement	source	date	target
0	Paul Specht	Mitchell County, North Carolina, sheriff's dep...	TikTok posts	January 6, 2025	false
1	Kwasi Gyamfi	"They still haven't cracked" Thomas Matthew Cr...	Instagram posts	• January 6,	false
2	Caleb McCullough	Photo shows a child "found" in several cities.	Facebook posts	January 6, 2025	false
3	Ciara O'Rourke	"22 states will not be certifying the (2024) e...	Threads posts	January 6, 2025	false
4	Madison Czopek	"BREAKING: A second attack in New Orleans has ...	Facebook posts	January 6, 2025	false

Figure 19.1 –Polifact data (Web Scraping)

- The LIAR2 dataset, a new benchmark dataset of around 23k manually labeled by professional fact-checkers for fake news detection tasks. Columns includes: 'id', 'label', 'statement', 'date', 'subject', 'speaker', 'speaker_description', 'state_info', 'true_counts', 'mostly_true_counts', 'half_true_counts', 'mostly_false_counts', 'false_counts', 'pants_on_fire_counts', 'context', 'justification'. In the image below there is too many columns to be shown.



	id	label	statement	date	subject	speaker	speaker_description	state_info	true_counts	mostly_true_counts	half
0	13847	5	90 percent of Americans "support universal bac...	October 2, 2017	government regulation;polls and public opinion...	chris abele	Chris Abele is Milwaukee County Executive, a p...	wisconsin	1	4	
1	13411	1	Last year was one of the deadliest years ever ...	May 19, 2017	after the fact;congress;criminal justice;histo...	thom tillis	Thom Tillis is a Republican who serves as U.S....	north carolina	0	2	
2	10882	0	Bernie Sanders's plan is "to raise your taxes ...	October 28, 2015	taxes	chris christie	Chris Christie announced June 6, 2023 that he ...	national	21	20	
3	20697	4	Voter ID is supported by an overwhelming major...	December 8, 2021	voter id laws	lee zeldin	Lee Zeldin is a Republican representing New Yo...	new york	1	2	
4	6095	2	Says Barack Obama "robbed Medicare (of) $716 b...	August 12, 2012	federal budget;history;medicare;retirement	mitt romney	Mitt Romney is a U.S. senator from Utah. He ra...	national	31	33	

Figure 19.2 – Hugging Face data

Final dataset.

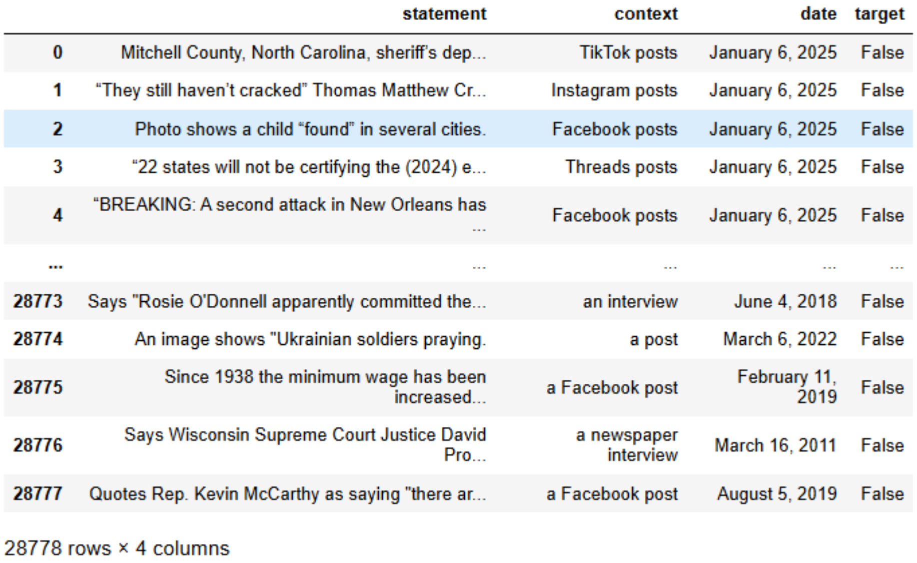
To derive the final dataset before data cleaning, the following steps are taken:

1. Derive the target column

2. Rename columns for consistency
3. Format the date column uniformly

First, to derive the target column, I define it as follows: statements that are at least half true are marked as "true," and the rest are categorized as "false" for fake news. Renaming the columns was straightforward. However, the date column required special attention due to inconsistencies, such as dots at the start of some dates and missing years. To resolve this, I removed the dots and used the next records to determine the correct year for each record. If this was not possible, I used the previous record's year as a fallback. The data period is 2000 to 2022.

After applying these changes, the resulting dataset has 28,778 rows and 4 columns: 'statement', 'context', 'date', and 'target'. The final dataset output is as follows:



	statement	context	date	target
0	Mitchell County, North Carolina, sheriff's dep...	TikTok posts	January 6, 2025	False
1	"They still haven't cracked" Thomas Matthew Cr...	Instagram posts	January 6, 2025	False
2	Photo shows a child "found" in several cities.	Facebook posts	January 6, 2025	False
3	"22 states will not be certifying the (2024) e...	Threads posts	January 6, 2025	False
4	"BREAKING: A second attack in New Orleans has ...	Facebook posts	January 6, 2025	False
...	...	...	...	...
28773	Says "Rosie O'Donnell apparently committed the...	an interview	June 4, 2018	False
28774	An image shows "Ukrainian soldiers praying.	a post	March 6, 2022	False
28775	Since 1938 the minimum wage has been increased...	a Facebook post	February 11, 2019	False
28776	Says Wisconsin Supreme Court Justice David Pro...	a newspaper interview	March 16, 2011	False
28777	Quotes Rep. Kevin McCarthy as saying "there ar...	a Facebook post	August 5, 2019	False

28778 rows × 4 columns

Figure 20 – Combined data

Exploratory Data Analysis

Afterwards, EDA and data cleaning was carried out. The following are the findings for more in-depth refer to the jupyter notebook:

Initial dataset

Timeline chart

We can see that there is more data in the recent year which is good as we do not want any data drift so keeping it from 2019 onwards will help to focus on the current trends.

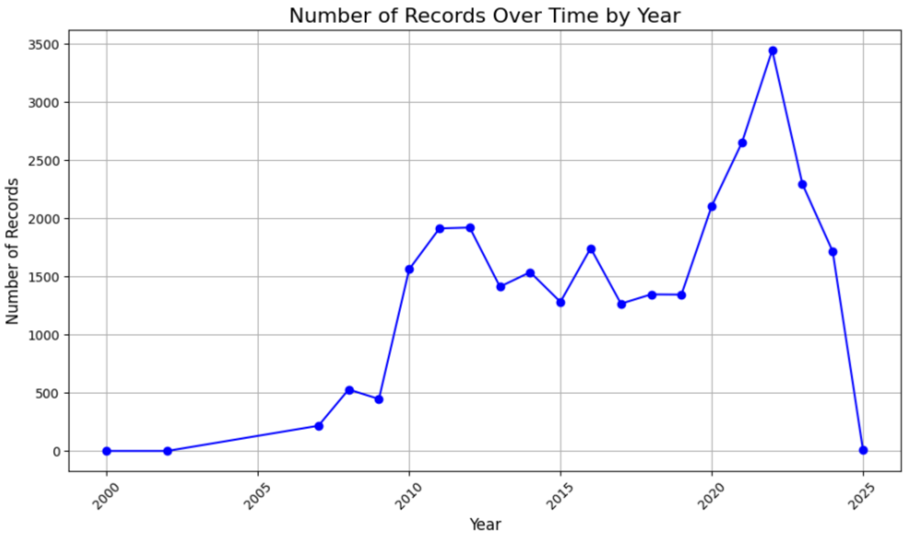


Figure 21 – Number of records over the years

Insights from EDA

Charts from the analysis are available in the individual submission here is a summary:

Aspect	Description
Most Common Words	The most frequent words (e.g., 'the,' 'and,' 'a') are stop words that carry little meaning and produce noise. These words increase dimensionality in TF-IDF and CountVectorizer, potentially causing longer model load times.
No Null Values in Dataset	Confirmed that the dataset does not contain any null values.
Punctuation Frequency	Punctuation does not add much meaningful information and increases dimensionality, creating noise for the model. Many text records are affected by punctuation, which needs to be addressed.
Presence of Mixed Words and numbers	Numbers provide little information and increase dimensionality. Mixed words (e.g., '1st, 6th, 50th ') combine numbers and letters, disrupting linguistic regularities. These words, along with standalone numbers, create inconsistencies and negatively impact the model. Exception is COVID19 which is a common topic as seen in the word cloud above. Thus, we will not be removing it.
Presence of Special Characters or HTML Tags	Special characters and HTML tags do not provide meaningful insights and create noise. Although most text does not have these characters, they still need to be removed to improve model performance.

Outlier Detection Based on Statement Length	Box-and-whisker plots of text length reveal several outliers outside the interquartile range. These extreme values prevent the model from generalizing well and negatively affect performance.
Inconsistent formatting	There are different word forms (e.g. said and says), these words have the same meaning and additionally some are capitalized. Thus, this could increase dimensionality and noise in the model training process. However, for contractions, we leave it in as it may have some meaning to it.
Loss of context	By treating the top 100 bigrams as compound words we can retain some of the context.
Imbalance dataset	5:1 ratio between Fake news and truthful news. This may skew the model to predict fake news.

Data Preparation

Time to clean using the following steps:

- No Null values
- Remove outliers based on sentence length
- Remove words with special characters or HTML tags
- Remove Stop words and punctuations
- Remove numbers and mixed words
- Convert all words to lower casing
- Get the top 100 bigrams and treat them as compound words and tokenize them together
- Lemmatize the words

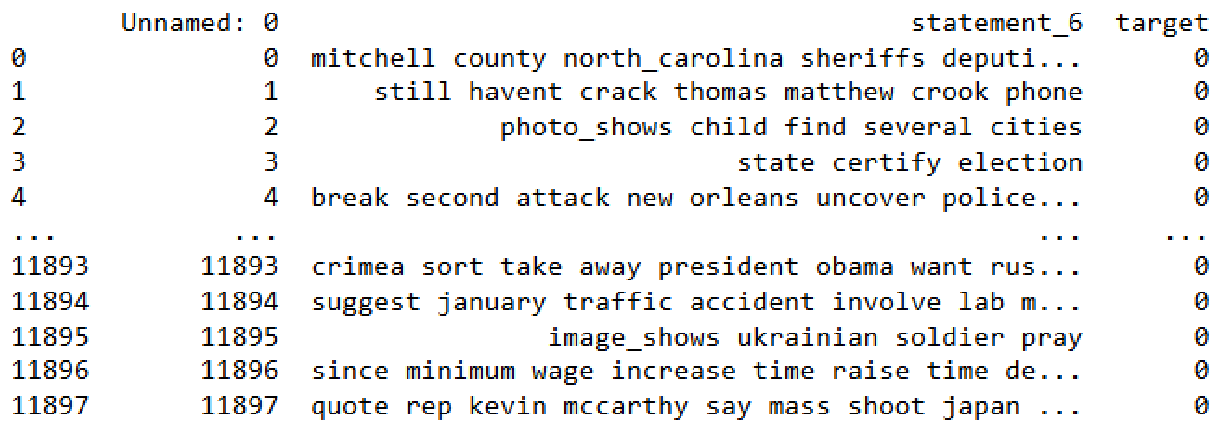
The final cleaned data has 11898 rows. Afterwards, three files were saved: one containing the list of top bigrams in text format, one with the fully cleaned dataset including all columns, and one with only the final cleaned statement column and the target column.

Filtered dataset

After filtering for recent years (2019 onwards), the dataset was reduced to 13,581 records. Following the data preparation steps, the dataset now contains approximately 12,000 records. Although smaller, this size remains sufficient for analysis.

The dataset includes two columns:

- statement_6: The cleaned text of the news statement.
- Target column: Labels where False represents fake news, and True represents 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
```

Figure 22 – Final Data

Data Preprocessing

Models cannot directly process raw text data, so it must first be transformed into numerical representations using vectorization techniques such as TF-IDF (Term Frequency-Inverse Document Frequency), Count Vectorizer, Word2Vec, or BERT. These methods convert text into a format that machine learning models can understand.

Class Imbalance solutions

To address class imbalance, we apply oversampling using SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic examples for the minority class by creating new data points based on the nearest neighbors, thus balancing the dataset. However, oversampling needs to be applied separately for each preprocessing technique (TF-IDF, Count Vectorizer, Word2Vec, BERT) to maintain consistency in the dataset.

Under sampling was not considered because our primary goal is to capture as much fake news (labeled as 0) as possible. The presence of more data points from the minority class (fake news) will help improve the model's ability to detect fake news. Reducing the number of fake news examples through under sampling could lead to a loss of valuable information needed for accurate predictions.

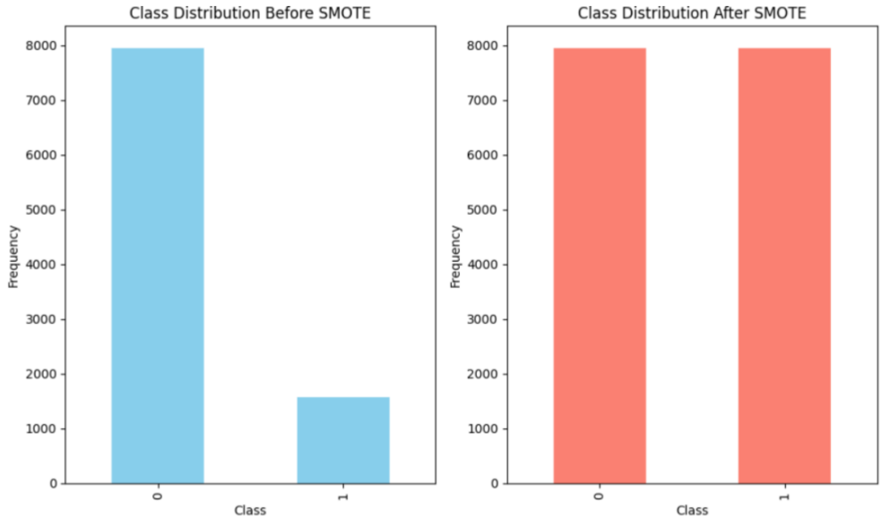


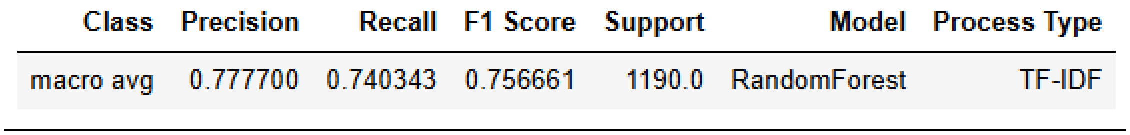
Figure 23.1 – Before and after SMOTE (Oversampling)

Another way to deal with the imbalance data is through the use of weighted class.

Data processing results

To determine the best data processing method among TF-IDF, Word2Vec, Count Vectorizer and BERT we will create a simple base model using Random Forest. Simultaneously, we will evaluate whether weighted class or oversampling yields better results. For simplicity in evaluation, we will filter by the macro average. The macro average accounts for class imbalances by considering each class having equal importance, providing a more balanced metric.

<u>Conclusion</u>



Class	Precision	Recall	F1 Score	Support	Model	Process Type
macro avg	0.777700	0.740343	0.756661	1190.0	RandomForest	TF-IDF

Figure 23.2 – Results of optimal combination of preprocessing

Upon comparing Recall, Precision, and F1 score, it appears that a simpler technique like TFIDF outperforms the more complex BERT and other techniques based on the F1 Score. Therefore, for model development moving forward, we will proceed with the TFIDF combined with Weighted class. For more please refer to the jupyter notebook.

Classification of Deepfakes Online (Wei Jun)

Solution

An image classification model that identifies whether an image is a deepfake or real image based on the face.

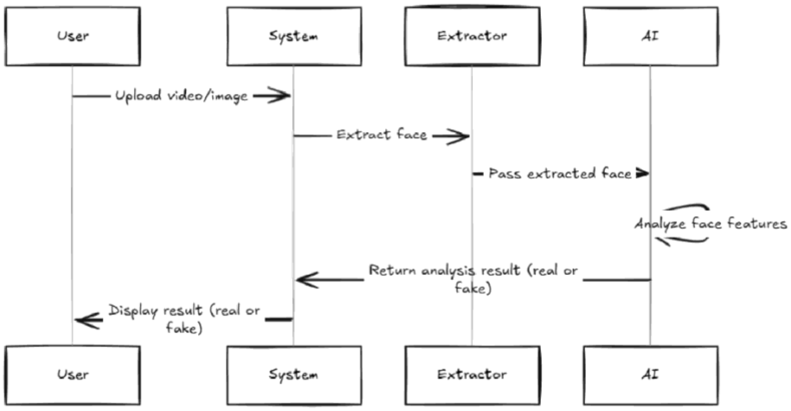


Figure 24 – System diagram for deepfake detection online

Core Model:

- An **image classification model** to detect whether an image is a deepfake or real based on facial features.
- Lightweight for fast identification
- Reliable performance for trustworthiness
- Experimenting with **training a CNN model training from scratch** and using **pre-trained models with transfer learning** to compare results.

Additional Feature:

- Utilize **Google's Face Detection API** to extract facial regions from videos.
- Pass extracted faces into the deepfake detection classifier for accurate predictions.

The model prediction is just a suggestion and should require human intervention to determine if video is actually real or fake.

Objectives of Project

The goal is to empower online users to protect themselves against deepfake scams by providing tools and awareness to identify and mitigate these threats. By tackling the misuse of deepfake technology, we aim to create a safer and more trustworthy digital environment, particularly on social media, where users can interact without fear of deception or fraud.

1. **Data Visualizations and Insights:** Perform Exploratory Data Analysis (EDA) on image datasets to uncover patterns, anomalies, and errors. Followed up by cleaning and preprocess image data to ensure consistency and quality.
2. **Experimentation with Models:** Explore various deep learning architectures, including Convolutional Neural Networks (CNNs) and pre-trained models with transfer learning. Followed up by performing hyperparameter tuning to optimize model performance and address challenges like overfitting.
3. **Performance Benchmarks:** The model's performance will be evaluated using Recall, Precision, and F1-Score, with a primary focus on Recall to minimize missed deepfakes and reduce misinformation risks. While Recall is prioritized, I am also targeting an F1-Score of 0.7 as it provides a balanced measure of both Precision and Recall, ensuring an overall effective assessment of model performance.
4. **Additional Feature:** As the Image classification model is trained to identify deep fakes based on faces only, we will need an external AI to extract and crop out the face from an image to pass it to the deep learning model for classification.

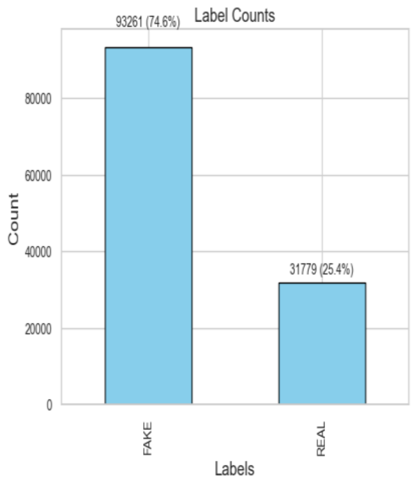
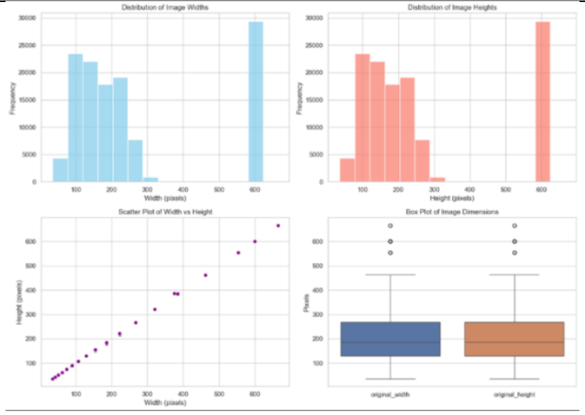
Dataset Used

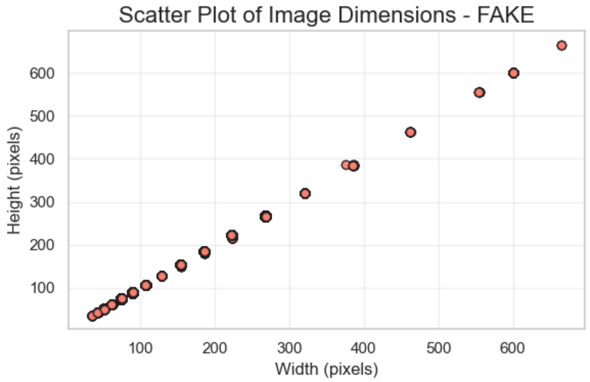
This dataset is taken from Kaggle, it takes the first frame from all videos in here, which has over 470GB worth of videos of real and deep fake videos. This dataset is a subset which contains over 95,000 images: Kaggle dataset

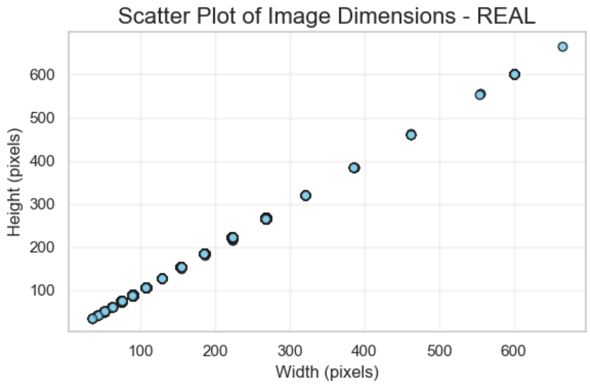
This dataset is taken from Roboflow. Its images have filters applied, (i.e 1 image has 4 versions with different filters) helping the model to be more robust. This dataset contains close to 30,000 images: Roboflow dataset

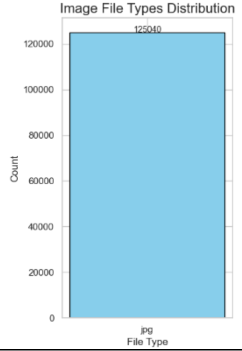
The combined dataset contains 2 classes. **Real** and **Fake**.

Exploratory Data Analysis

Image	Description
	This shows the distribution of labels between Real and Fake. There is close to 3 times more Fake (74.6%) than Real (25.4%) images
	The above image shows the distribution of the image widths and heights. The widths and heights are distributed quite sparsely. however, there is a large amount at 600 height and 600 width. small amounts fall below 100 width and height. Majority of the images are square, (i.e equal width and height) there are some that width is larger than height. Both width and height have a mean of about 180px. Box plots of width and height are very identical

<table>
<tr>
<td>




</td>
<td>Shows the comparison between image dimensions of real and fake class. Fake chart has a sligth difference in image size at 390 height and 380 width.
Not all images are squares.</td>
</tr>
<tr>
<td>


</td>
<td>All images are of JPG type.</td>
</tr>
</table>

Data Preparation

Steps done

1. **Adjusted all image sizes to 224 x 224**
2. **Split data into 2 folders, Real and Fake**
3. **Remove duplicate images**
4. **Balance class (under sampling as there is sufficient real images)**

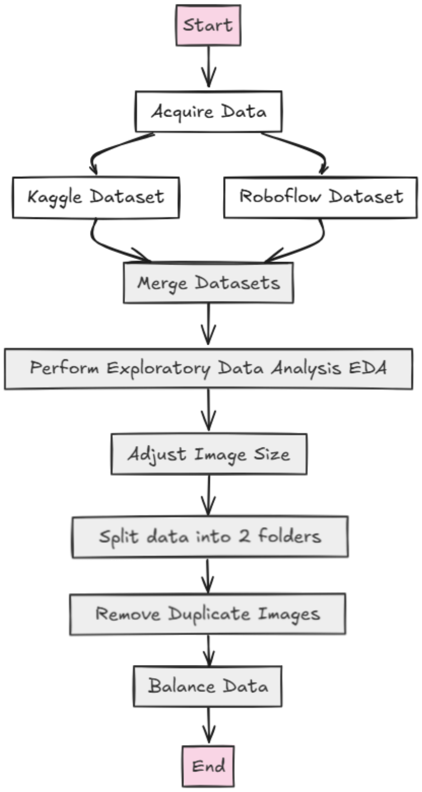


Figure 25 – Data processing pipeline

Example data

Real:

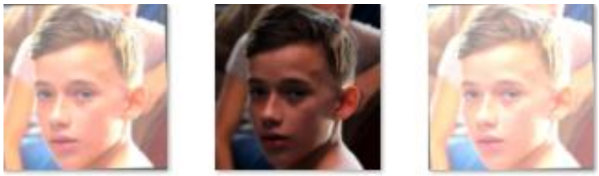


Figure 26 – Example of real image data

Fake:

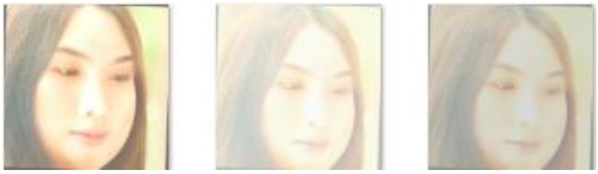


Figure 27 – Example of fake image data

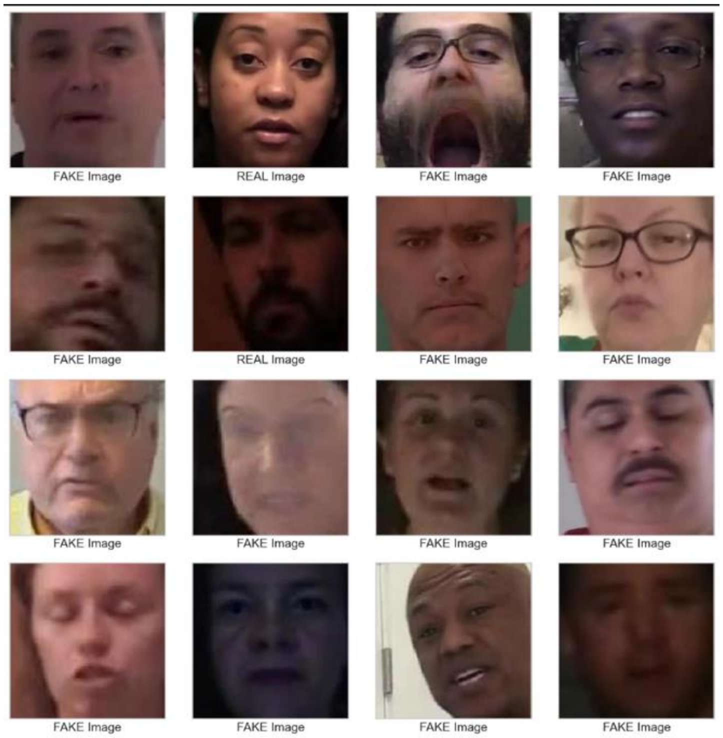


Figure 28 – Random sample of images

Classification of Mental Health Based on Social Media Text (Pin Shien)

Solution

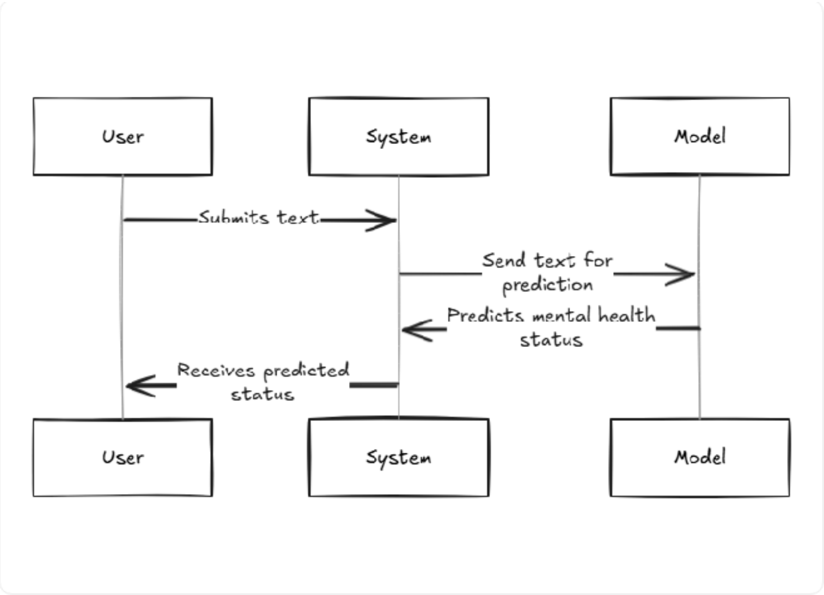


Figure 29 – System diagram of mental health based on social media

Objectives Of Project

The primary objective is to develop a text classification model on social media posts, utilizing textual data from platforms (e.g., Twitter, Reddit, etc). The classifier will then determine whether a given post falls under one of several mental health labels (e.g., Normal, Depression, Anxiety, Stress, Suicidal, etc.). To achieve this, the following steps are needed:

1. **Data Visualizations and Insights:** Both user-generated text and any available metadata will be visualized and cleaned to identify anomalies or errors. Exploratory Data Analysis (EDA) will be conducted to understand class distributions, common keywords, and potential imbalances.
2. **Experimentation with Models:** Various embedding approaches will be tested alongside classical machine learning models and deep learning models to see which generalizes best to the problem. Hyperparameters will be fine-tuned to achieve optimal model performance.
3. **Performance benchmarks:** The model will use F1-score or to measure how effectively the model separates each mental health category, especially under class imbalance. A higher F1-score indicates better balance between precision and recall, making it ideal for comparing different models and approaches.
4. **Additional Feature:** Due to the sensitive nature of mental health classification, explainability techniques (e.g. SHAP) will be used to clarify why a post is tagged with a specific label. Pretrained models with transfer learning will be leveraged to enhance model accuracy, efficiency, and contextual understanding. These models will be fine-tuned on the dataset to better capture the nuances of mental health expressions in text.

Dataset Used

Link to Dataset: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data

This dataset is designed for sentiment analysis in the context of mental health. It provides an excellent resource for training machine learning models to analyze and predict mental health conditions based on textual data. The data is collected from a variety of sources, including social media posts, Reddit, Twitter, and more, offering a diverse range of expressions and contexts. Each entry is labeled with a specific mental health category, making it a valuable tool for conducting detailed sentiment analysis.

Data Overview:

- The dataset contains over 53,000 entries, each comprising a text statement and a corresponding mental health status.
- The labels span seven mental health categories: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder.
- This dataset serves as a comprehensive resource for exploring the relationship between language use and mental health

Exploratory Data Analysis

Countplot

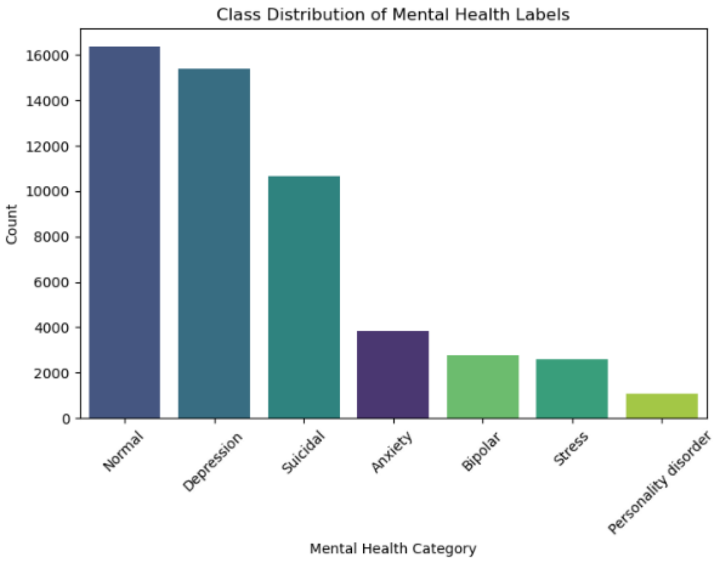


Figure 30 – class distribution of mental health labels

The initial exploratory data analysis focused on understanding the distribution of the . A countplot was generated to visualize the frequency of each mental health category within the dataset. The plot revealed a significant class imbalance, with "Normal" being the most frequent category, followed by "Depression" and "Suicidal." The categories "Anxiety," "Bipolar," "Stress," and "Personality Disorder" exhibited considerably lower frequencies. This class imbalance will be an important consideration during model development and evaluation, as it can potentially bias the model towards the majority class.

Histogram

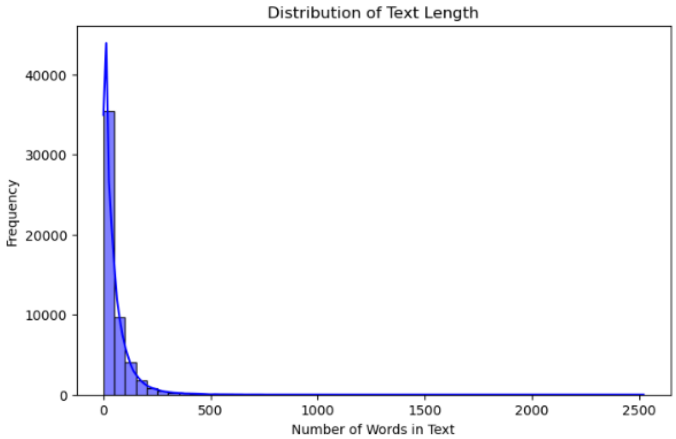


Figure 31 – distribution of text length

The exploratory data analysis delved into the distribution of text lengths within the dataset. A histogram was generated to visualize the frequency of text lengths, measured by the number of words. The plot revealed a right-skewed distribution, indicating that most of the texts are relatively short, with a long tail suggesting the presence of some very long texts. This observation suggests that text length might be a relevant feature for further analysis and could potentially impact model performance. The presence of a long tail might necessitate careful consideration of preprocessing techniques such as truncation, padding, or feature scaling to mitigate the influence of outliers and ensure that the model is not unduly influenced by the very long texts.

Before: Barplot

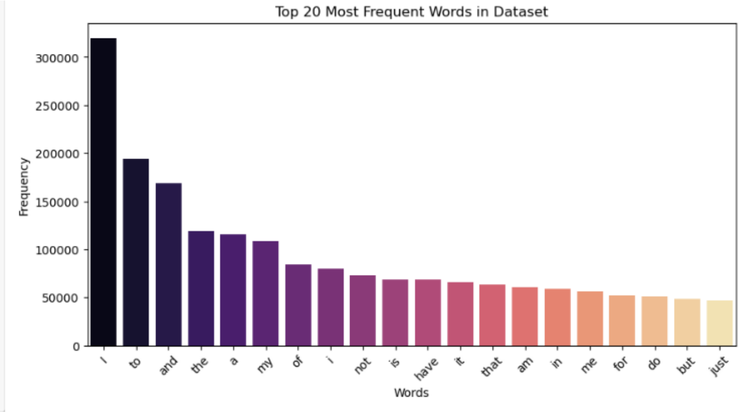


Figure 32 – most frequent words in dataset

The exploratory data analysis began with an examination of the most frequent words in the dataset. A bar plot was generated to visualize the top 20 words, revealing that the dataset is dominated by common English words such as "the," "to," "and," "a," and "of." These words, often referred to as stopwords, typically carry little semantic meaning and can be removed during preprocessing to improve the focus on more informative words.

After: Barplot

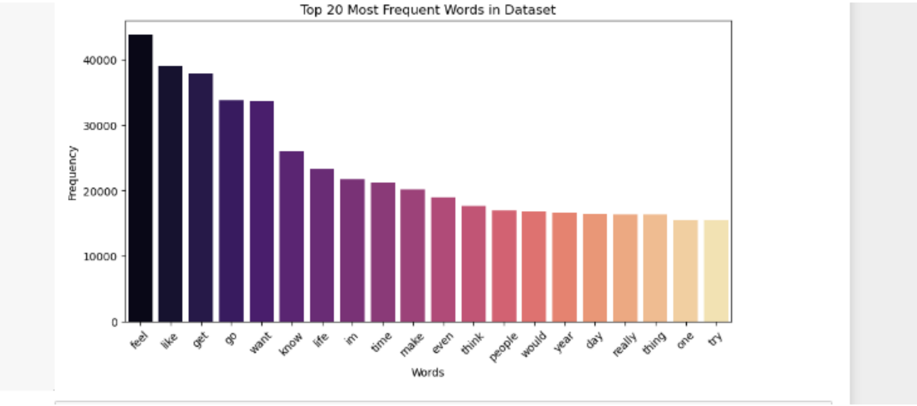


Figure 33 – most frequent words in dataset after processing

Following the removal of stop words and lemmatization, a renewed analysis of the most frequent words was conducted. The bar plot revealed a shift in the distribution, with words like "feel," "like," "get," and "want" now appearing among the most frequent. This indicates that the preprocessing steps effectively removed common English words, allowing for a clearer focus on more meaningful terms. The presence of these words suggests that the text data may be related to personal emotions, experiences, and desires.

Word Cloud

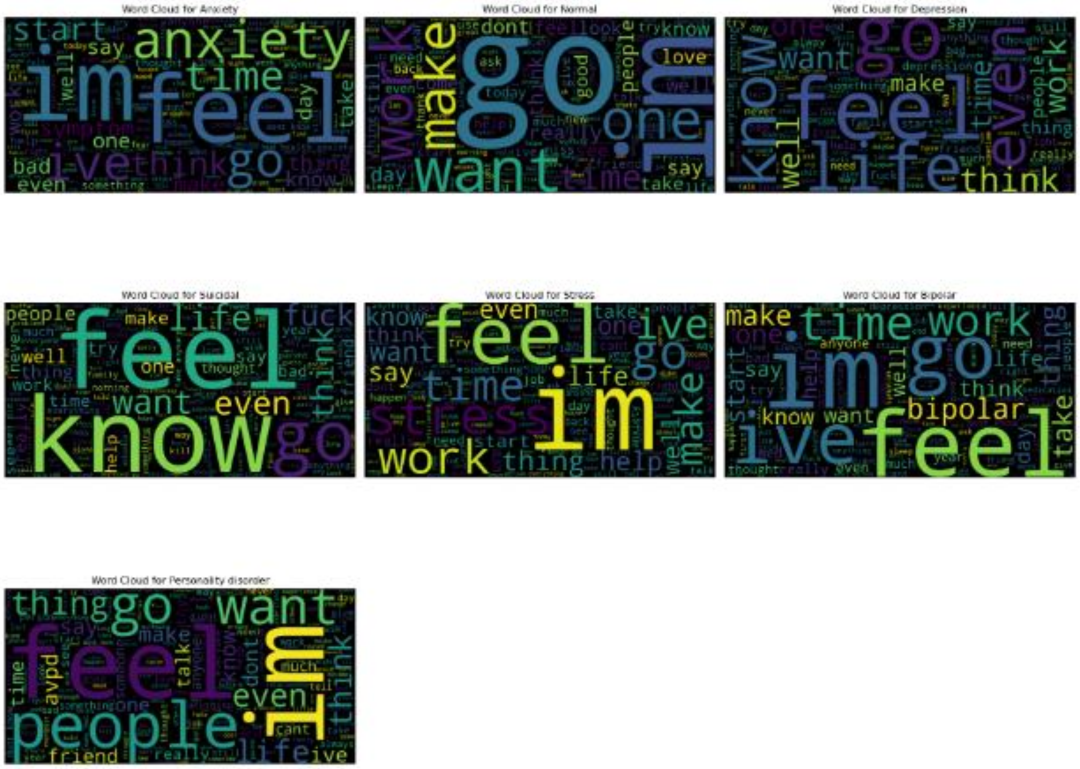
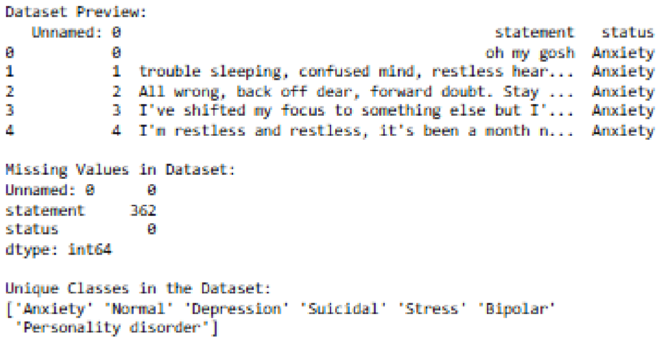


Figure 34 – Word cloud

The word cloud provides a visual representation of the most frequently used words for each mental health status category, offering insights into the linguistic patterns associated with different conditions. The dominant presence of the word "feel" across all categories highlights the emotional nature of these statements, suggesting that self-expression through feelings is central to discussions of mental health. Categories like *Anxiety* prominently feature words like "anxiety," "time," and "think," reflecting common themes of overthinking and time-related stress. *Normal* includes terms such as "go," "work," and "make," indicative of everyday activities and goals. *Depression* and *Suicidal* are characterized by words like "feel," "know," and "life," emphasizing introspection and emotional struggle. Similarly, *Stress* and *Bipolar* share terms such as "I'm," "work," and "make," pointing to concerns about performance and coping mechanisms. Finally, *Personality Disorder* includes terms like "people" and "want," possibly highlighting social and interpersonal challenges. This analysis underscores how language varies across mental health statuses and can aid in understanding the unique concerns of individuals within these categories.

Data Preparation



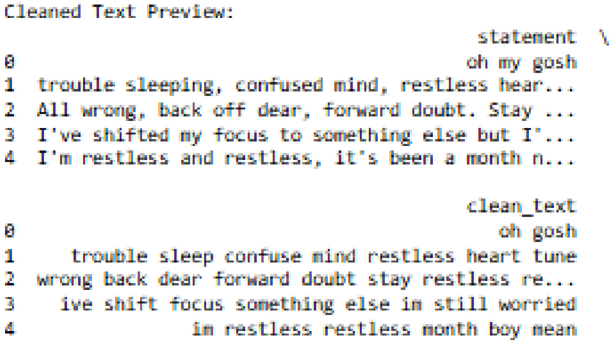
```
Dataset Preview:
   Unnamed: 0                                          statement   status
0           0                                         oh my gosh  Anxiety
1           1  trouble sleeping, confused mind, restless hear...  Anxiety
2           2  All wrong, back off dear, forward doubt. Stay ...  Anxiety
3           3  I've shifted my focus to something else but I'...  Anxiety
4           4  I'm restless and restless, it's been a month n...  Anxiety

Missing Values in Dataset:
Unnamed: 0      0
statement     362
status          0
dtype: int64

Unique Classes in the Dataset:
['Anxiety' 'Normal' 'Depression' 'Suicidal' 'Stress' 'Bipolar'
 'Personality disorder']
```

Figure 35 – dataset preview

The dataset preparation process involved several crucial steps to ensure data quality and readiness for analysis. Initially, the dataset was loaded using the pandas library, which provided an overview of its structure, including column names, data types, and sample rows. A check for missing values revealed that the "statement" column had 362 missing entries. These rows were subsequently dropped to maintain the consistency and integrity of the dataset. Additionally, the "status" column, which serves as the target variable, was analyzed to identify the unique mental health categories present. It was confirmed that the dataset includes seven distinct statuses: Anxiety, Normal, Depression, Suicidal, Stress, Bipolar, and Personality Disorder. After cleaning and verifying the data, the dataset was finalized, ensuring that it was free of missing values and properly structured for further exploratory data analysis (EDA) and preprocessing tasks.



```
Cleaned Text Preview:
                                           statement  \
0                                         oh my gosh   
1  trouble sleeping, confused mind, restless hear...   
2  All wrong, back off dear, forward doubt. Stay ...   
3  I've shifted my focus to something else but I'...   
4  I'm restless and restless, it's been a month n...   

                                          clean_text  
0                                            oh gosh  
1     trouble sleep confuse mind restless heart tune  
2  wrong back dear forward doubt stay restless re...  
3   ive shift focus something else im still worried  
4                 im restless restless month boy mean  
```

Figure 36 – cleaned text preview

To prepare the textual data for analysis, a text-cleaning pipeline was implemented using the Natural Language Toolkit (NLTK) library. First, necessary resources such as tokenizers, stop word lists, and part-of-speech (POS) taggers were downloaded. A custom function was developed to enhance the lemmatization process by incorporating POS tagging, ensuring more accurate reduction of words to their root forms. The cleaning process involved multiple steps: converting all text to lowercase, removing extra spaces, eliminating punctuation and numerical values, tokenizing the text into individual words, and filtering out common stopwords using NLTK's predefined list. Finally, each word was lemmatized based on its POS tag to improve semantic representation. The cleaned text was then added as a new column in the dataset, making it ready for exploratory analysis and modeling. A preview of the cleaned text confirmed the successful application of the pipeline, with statements transformed into simpler and more meaningful representations.

Milestone and Deliverables

Each member will be responsible to complete the tasks for their sub-problem using the CRISP-DM framework which includes:

- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment
- Documentation of all steps

On top of that, each member will also have unique responsibilities to ensure fair and efficient task distribution:

Team Member	Individual Tasks
Karthik	**Working on Harmful Meme Classification** • Help to oversee and review the project • Communication with the relevant stakeholders • Work on the background information and delegation of tasks/responsibility 1. **Data Preparation:** a. Perform Exploratory Data Analysis (EDA) to understand dataset characteristics. b. Balance the dataset using web scraping c. Extract the textual data from the dataset using an OCR d. Clean the textual and image data using python 2. **Model Development:** a. Train different neural network models with different parameters. b. Compare model performance using metrics like AUROC, accuracy and f1-score. c. Experiment with different hyperparameters and save the best-performing model. 3. **Integration:** a. Load the final model onto a web application for easy inferencing.

	b. Use vision language models to explain the harmful component in the memes. 4. **Testing and Validation:** a. Test model robustness with diverse datasets. b. Evaluate edge cases (e.g. low-resolution images). 5. **Deployment:** a. Ensure smooth integration into the webpage. b. Provide documentation for usage and interpretation of results.
Wei Jun	**Working on DeepFake Classifier** • Work on the timeline of the project • Determine tools and libraries used for the project 1. **Data Preparation:** a. Merge datasets from Kaggle and Roboflow. b. Perform **Exploratory Data Analysis (EDA)** to understand dataset characteristics. c. Clean and balance the data. 2. **Model Development:** a. Train different deep learning models (i.e., CNN, transfer learning models). b. Compare model performance using metrics like accuracy, precision, recall and f1 score. c. Tune hyperparameters and save the best-performing model. 3. **Integration:** a. Load the final model onto a **web application** for real-time predictions. b. Integrate **Google Face Detection API** for face extraction from videos. 4. **Testing and Validation:** a. Test model robustness with diverse datasets. b. Evaluate edge cases (e.g., low-resolution images, occluded faces). 5. **Deployment:** a. Ensure smooth integration into the webpage. b. Provide documentation for usage and interpretation of results.
Jun Ming	**Working on Classification of Fake News Online using social media text** • Work on the business scenario of the project • Identify the project risks and solutions to mitigate them 1. **Data Preparation:**

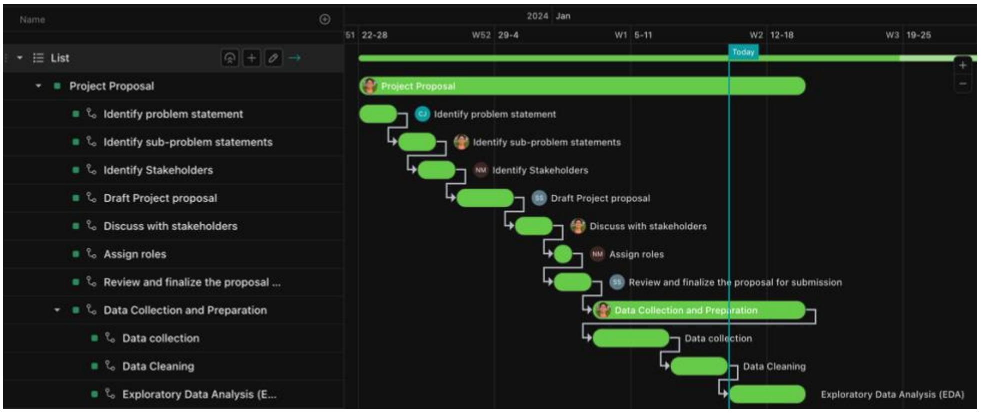
	a. Merge datasets from hugging face and web scraping websites b. Combine the datasets and format them in a standardized format. Perform EDA to get insights and based on those insights clean the data accordingly. Process the data using techniques like BERT, Word2Vec, TFIDF and count vectorizer. Pair this with Oversampling and weighted class to deal with data imbalance. 2. **Model Development:** a. Train different models, including classical machine learning models (Logistic Regression, SVM), deep learning models (LSTM, CNN), ensemble learning methods (Random forest, gradient boosting). b. Compare model performance using metrics like accuracy, precision, recall and f1 score. c. Tune hyperparameters and save the best-performing model. 3. **Integration:** a. Load the final model onto a web application for real-time predictions. b. Integrate API that fetches supporting articles. 4. **Testing and Validation:** a. Test on testing set and validate on validation after fine tuning b. Evaluate edge cases (e.g., short text, languages other than English, Long text). 5. **Deployment:** a. Ensure smooth integration into the webpage. b. Provide documentation for usage and interpretation of results.
Pin Shien	**Working on classification of mental health using social media text** • Work on the introduction of the project • Work on the timeline of the project **1. Data Preparation:** a. Merge datasets from Kaggle. b. Clean and preprocess the data. c. Perform Exploratory Data Analysis (EDA) including class distributions, common keywords, and linguistic patterns in different mental health categories. **2. Model Development:** a. Train different models, including classical machine learning models (Logistic Regression, SVM) and deep learning models (LSTM, CNN). b. Fine-tune pretrained transformer models with transfer learning, such as BERT, to enhance contextual understanding.

	c. Compare model performance using metrics like accuracy, precision, recall and f1 score. d. Tune hyperparameters and save the best-performing model for deployment. **3. Integration:** a. Deploy the final model into a web application or API for real-time classification of mental health status based on user-submitted text. **4. Testing and Validation:** a. Test model robustness across different datasets to ensure generalizability. b. Evaluate performance on edge cases, including short text inputs, ambiguous statements, and sarcastic expressions. **5. Deployment:** a. Ensure smooth integration into the webpage. b. Provide documentation for usage and interpretation of results. c. Implement Explainable AI (XAI) techniques to help users and mental health professionals understand predictions and increase trust in AI decisions.

Milestones

- Week 9 - 13: Data Collection and Preprocessing
- Week 14 - 17: Model Development and Integration into Web Application
- Week 17 - 19: Final Project Checklist

Week 9 – 13 Milestone: Data Collection and Preprocessing



Name

List

Project Proposal

Identify problem statement

Identify sub-problem statements

Identify Stakeholders

Draft Project proposal

Discuss with stakeholders

Assign roles

Review and finalize the proposal ...

Data Collection and Preparation

Data collection

Data Cleaning

Exploratory Data Analysis (E...

2024 Jan
51 22-28
W52 29-4
W1 5-11
W2 12-18
W3 19-25
Today
Project Proposal
CJ Identify problem statement
Identify sub-problem statements
NM Identify Stakeholders
SS Draft Project proposal
Discuss with stakeholders
NM Assign roles
SS Review and finalize the proposal for submission
Data Collection and Preparation
Data collection
Data Cleaning
Exploratory Data Analysis (EDA)

Figure 37 – data collection and preprocessing

Week 14-17 Milestone: Model Development and Integration into Web Application

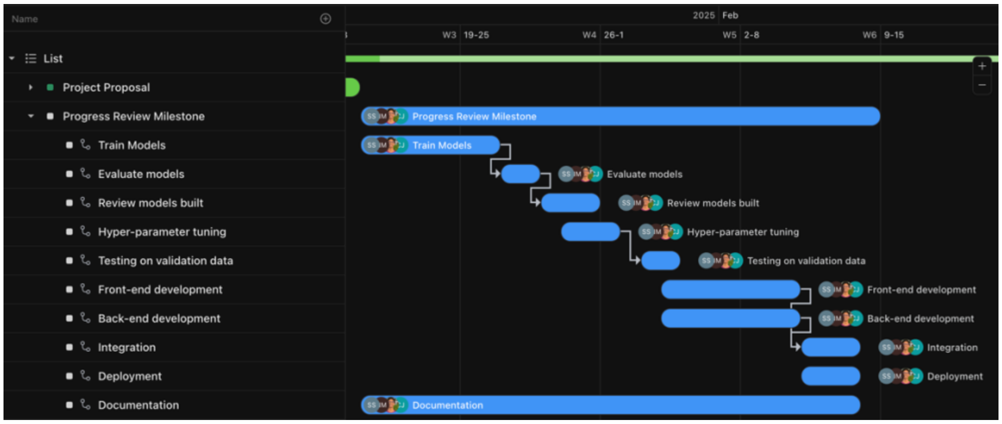


Figure 38 – model development and integration

**Note: when reviewing the model built, the additional features will also be implemented*

Week 19 Milestone: Final Project Checklist

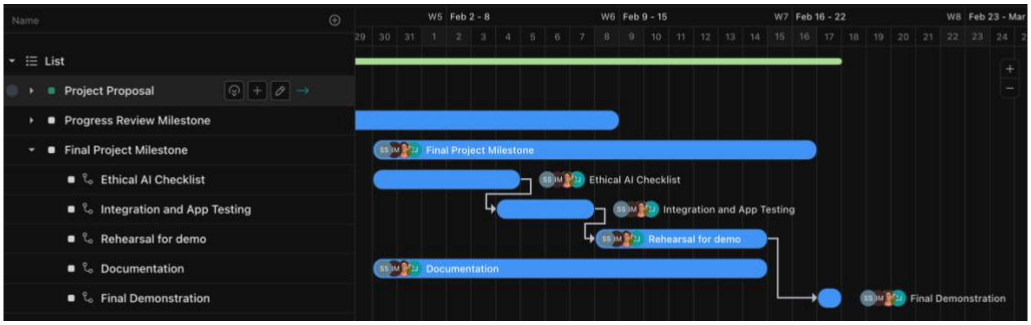


Figure 39 – Final project checklist

Tools used

Data Understanding and Preparation

- **Visualization:** Matplotlib and Seaborn
- **Tabular Data Analysis:** Numpy and Pandas
- **Text Processing:** NLTK, Gensim, and HuggingFace Transformers
- **Image Processing:** OpenCV, PIL, SciPy

Modelling and Hyperparameter Tuning

- **Deep learning library:** TensorFlow
- **Machine learning and hyperparameter tuning library:** Sci-kit learn, XG Boost

Deployment

- **Web server:** Flask / Fast API
- **Deployment:** AWS, Vercel, Google Cloud Platform, Azure (Based on RAM, compute and GPU requirements along with budget constraints)

Project Management

- **Code collaboration and versioning software:** GitHub
- **Project Management Tool**: Click Up

Generative AI

- **LLM:** GPT4 API and Gemini API
- **VLM:** Pali gemma
- **Miscellaneous APIs:** face detection API, web search API, entity detection API, etc.

Potential Risks and Ethical Considerations

In this section, the potential ethical considerations for each of the projects will be discussed along with ways on possible solutions.

1. **Hateful Meme Classifier**

    a. Develop the classifier using explainable AI techniques that allow stakeholders to understand how decisions are made. This transparency is essential for building trust, enabling accountability, and allowing for the identification and correction of potential biases.

    b. Approach the problem with sensitivity since some of the data that the model is using is sensitive to certain demographics. Ensure polite and considerate use of language so as to not offend anyone. For the examples, used as well, the less offensive memes were considered.

2. **Social Media Text Classifier**

    a. Misclassifying mental health can lead to harmful misinterpretations and stigmatization. To mitigate this, models must be validated using diverse data to minimize bias and improve accuracy. Importantly, these models should be presented as supportive tools within a comprehensive assessment, rather than definitive diagnoses.

    b. Social media data frequently contains sensitive personal information, such as personal opinions, political views, and intimate details of users' lives. Utilizing this data without explicit and informed consent constitutes a serious violation of privacy rights. To ensure ethical data usage, publicly available datasets with clear permissions are used for the project purposes.

3. **Bias Mitigation and Data Integrity**

    a. Proactively address potential biases in the dataset by ensuring it represents diverse user demographics and linguistic variations. Regularly monitor for data drifts which are shifts in the data distribution over time that could impact the model's accuracy and update the dataset and model as needed to sustain fairness and reliability.

4. **Privacy and Compliance**

    a. In compliance with PDPA (Personal Data Protection Act), limit data collection to the text content and the predicted classification. Avoid gathering unnecessary personal information to protect user privacy. Ensure all data processing aligns with legal and ethical standards.

5. **Continuous Improvement**

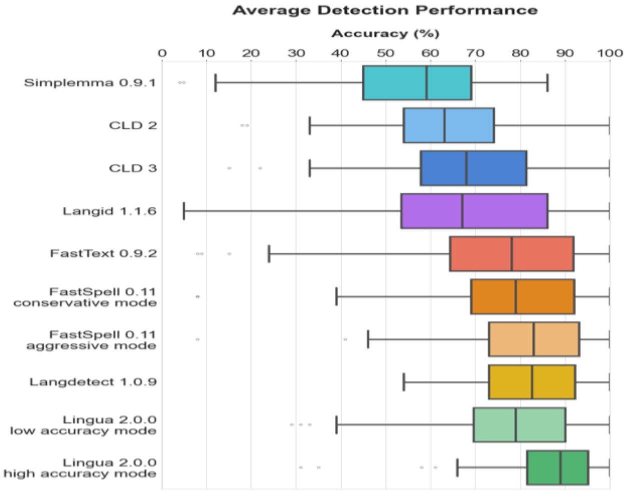
    a. Commit to an iterative improvement process by refining the model based on real-world performance and feedback. Incorporate mechanisms for feedback and periodic audits to detect and address ethical concerns or biases proactively.

Appendix

1. https://www.enterpriseappstoday.com/stats/memes-statistics.html
2. https://www.rsis.edu.sg/rsis-publication/rsis/deciphering-the-language-of-internet-memes-its-use-in-disinformation-and-detection-via-ai/
3. https://mothership.sg/2018/09/man-report-police-nyjc-meme-reason/
4. https://www.straitstimes.com/life/entertainment/in-bad-taste-national-crime-prevention-council-removes-amber-heard-meme
5. https://medium.com/@beyond_verse/how-social-media-impacts-mental-health-a-psychological-analysis-4b58f1c49283#:~:text=Affiliation%20with%20these%20communities%20can,health%20in%20the%20digital%20age.
6. https://www.channelnewsasia.com/singapore/rise-harmful-social-media-content-increase-those-inciting-racial-religious-tension-violence-online-safety-poll-4496021
7. https://www.straitstimes.com/singapore/sm-lee-warns-that-video-of-him-promoting-an-investment-scam-on-social-media-is-a-deepfake
8. https://www.police.gov.sg/-/media/4B5A6A81EDC4470EA8ED2B8CDE89EE4D.ashx
9. https://www.security.org/resources/deepfake-statistics/
10. https://documents1.worldbank.org/curated/en/099741206152335619/pdf/IDU07170b25f089b00406f0b6a40c67d0a1670b5.pdf
11. https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024.pdf
12. https://opengovasia.com/2024/01/22/singapores-proactive-approach-to-online-trust-and-safety/
13. https://www.nytimes.com/2020/06/27/technology/pizzagate-justin-bieber-qanon-tiktok.html
14. https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic

15. https://www.mddi.gov.sg/media-centre/press-releases/survey-by-mci-on-harmful-online-content-encountered-by-sg-users/

16. Image 1 – *Comparison the language detectors, lingua 2.0 on high accuracy mode was used for the meme language detection.*



17. https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2329610#d1e2461