Project Proposal

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Detecting signs of Mental Health Crises in Malaysian Social Media Text using Emotion Classification and Explainable AI

Choong Zi Xuan (MCS241038)

Video Presentation link:
https://youtu.be/0wjcHFqkysA

Introduction

Problem Background

- Mental health issues are rising in Malaysia
- Many users express emotional distress on social media like Reddit
- Mental health crises detection from social media text
- Current Study more is for sentiment analysis
- Lack of study for Malaysia data

Project Aim:

To flag out mental health crises by done the emotion classification for social media data while using Explanation Artificial Intelligence (XAI) to enhance the transparency and trustworthy of the predictions

Problem Statement

(a)Social media posts contain early signs of mental health crises such as depression but these signals are not systematically analysed for intervention.

(b)The prediction of machine learning is not transparent and let people cannot trust the prediction outcome.

(c)Existing research lacks culturally and linguistically tailored models for Malaysia populations. No previous studies have use local datasets for crises prediction.

Research Objective

1. To collect and preprocess mental health related social media data from Malaysian Reddit users for emotion-based analysis.
2. To apply a pretrained DistilBERT emotion classification model to detect high-risk emotional states in Malaysian social media text.
3. To interpret the model predictions using Explainable AI (XAI) techniques.

Scope

1. Data Sources: This study will do the web scrapping to get the real time dataset from Reddit.
2. Target Conditions: This study aim to get the outcome of emotion classification with flag out the post with emotion that have high-risk for mental health crises.
3. Technical Focus: This study will use DistilBERT model to do the prediction of mental health crises and LIME to interpret the machine learning.

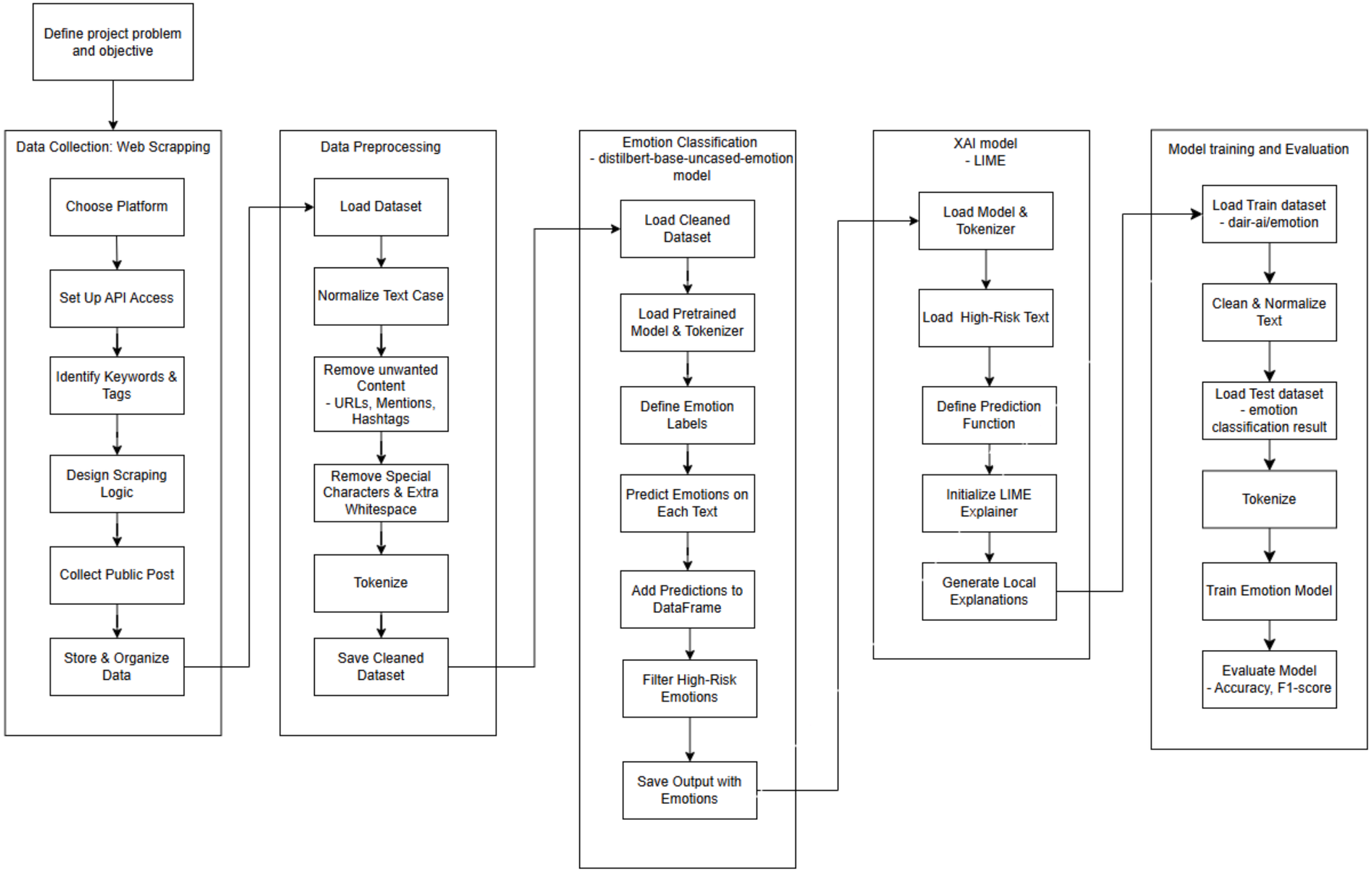
Literature Review

No	Title	Author	Dataset	Model Used	Result	Limitation
1	Machine learning Driven Analysis of Mental Health Indicators in social media Posts	(Garg, Garg, Dixit, & Pandey, 2024)	Reddit	Decision Tree, Random Forest, BERT	BERT have the highest result for accuracy, precision, recall and F1-score. 0.82 for all.	• Lack of evidence in utilizing the most active web people for obtaining the most accuracy results • The study focuses on Reddit, limit the generalizability of the findings to other social media platforms or demographics
2	Predicting Mental Health Disorder on Twitter Using Machine Learning Techniques	(Lim, Kamarudin, Ismail, Ismail, & Kamal, 2023)	Twitter	SVM, Decision Tree, Naïve Bayes	SVM have the highest accuracy.	• Just focus on the data on Twitter • The dataset used limit and small
3	Mental illness detection using sentiment analysis in social media	(Odja, Widiarta, Purwanto, & Ario, 2024)	Reddit	KNN, Random Forest, Neural Network	Random Forest have the best performance with 80.6% for F1-score, accuracy, recall and precision.	• The dataset is small that only consist 350 columns of data.

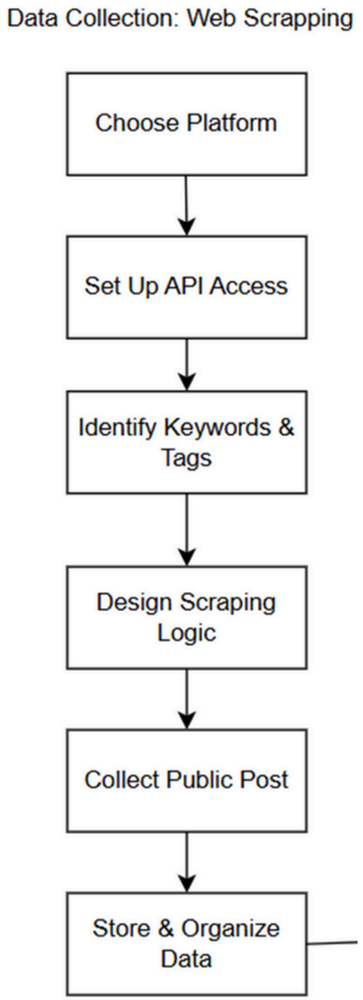
No	Title	Author	Dataset	Model Used	Result	Limitation
4	Lightweight advanced deep learning models for stress detection on social media Posts	(Qorich & Ouazzani, 2025)	Reddit, Twitter	Lightweight deep learning methods, BERT, CNN	BERT achieved 85.67% of accuracy on small Reddit dataset; CNN reached 97.62% accuracy on Large Twitter dataset.	• Performance varied across platform • Platform-specific tuning required
5	Integrating Machine Learning and Sentiment Analysis: A Comparative Study on Mental Health Classification from Social Media Data	(Kaushik & Sharma, 2024)	Reddit, Twitter, Kaggle	Decision Tree, Logistic Regression, XGBoost	XGBoost have the highest accuracy (82%), logistic regression has 78% of accuracy and decision tree have 67% of accuracy.	• There are some mistakes in classification “Stress” and “Personality Disorder” even it has the highest accuracy.

Research Methodology

Research Framework



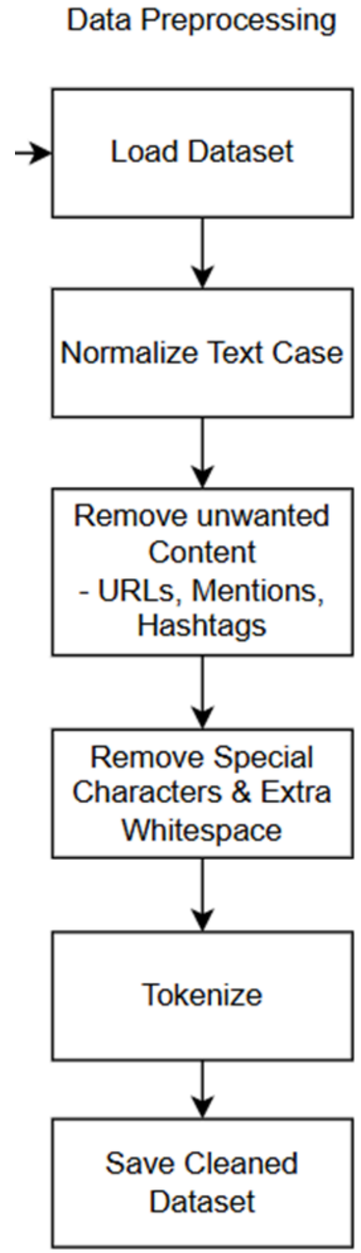
Data Collection : Web Scrapping

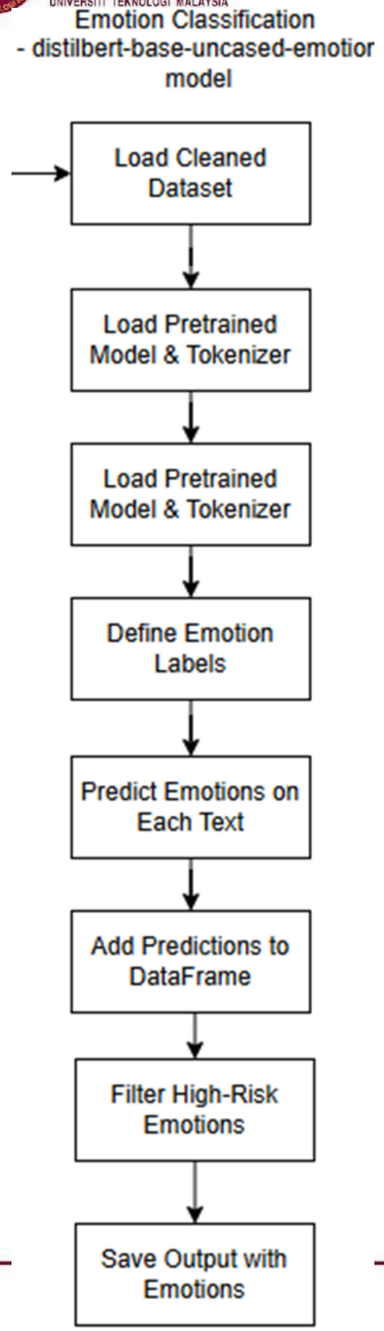


1. Use PRAW to scrape the post from Reddit.
2. Set the keywords for web scrapping: mental health, stress, depression, anxiety.
3. Set what to collect: post title, post body, comments, time posting and time of comment.
4. Start the web scraping.
5. Save the scraping post to csv file

Data Pre-Processing

1. Load the data that scrape from Reddit
2. Normalize the text
3. Remove unwanted content
4. Remove special characters
5. Tokenize the text
6. Save to a new column





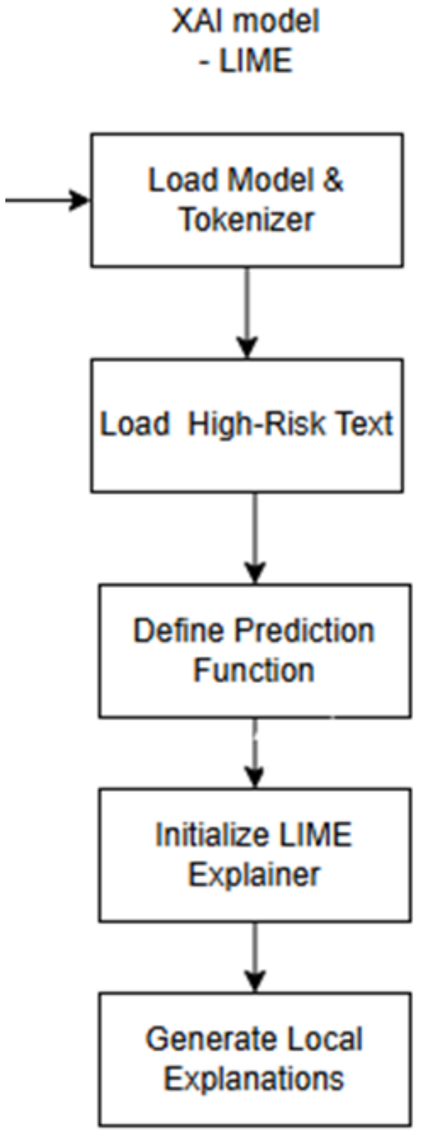
Emotion Classification

Model used: distilbert model
Model name: bhadresh-savani/distilbert-base-uncased-emotion

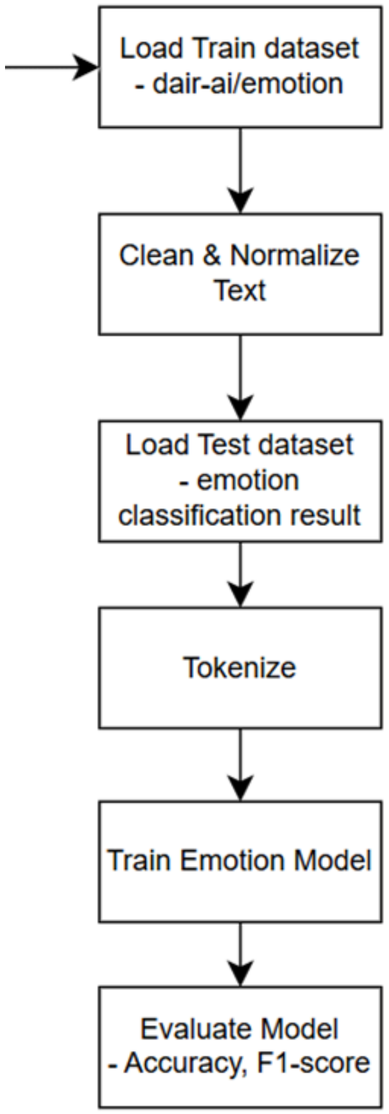
This model will classify the text data into 6 emotions:
sadness, joy, anger, love, fear, and surprise

XAI model interpretation

XAI model used: LIME
LIME will interpret what word have key contribute to the emotions prediction



Model training and Evaluation



Used the publicly available dataset ‘dair-ai/emotion’ as the training set of the model
The Reddit pseudo-labeled dataset as the test set.

Fine-tuning using the Hugging Face Trainer API.

Evaluation on :
Accuracy: to determine the overall proportion of correctly predicted emotion labels out of all predictions made
F1-score: to account for class imbalance and to ensure that each emotion category contributed equally to the final score

Initial Result

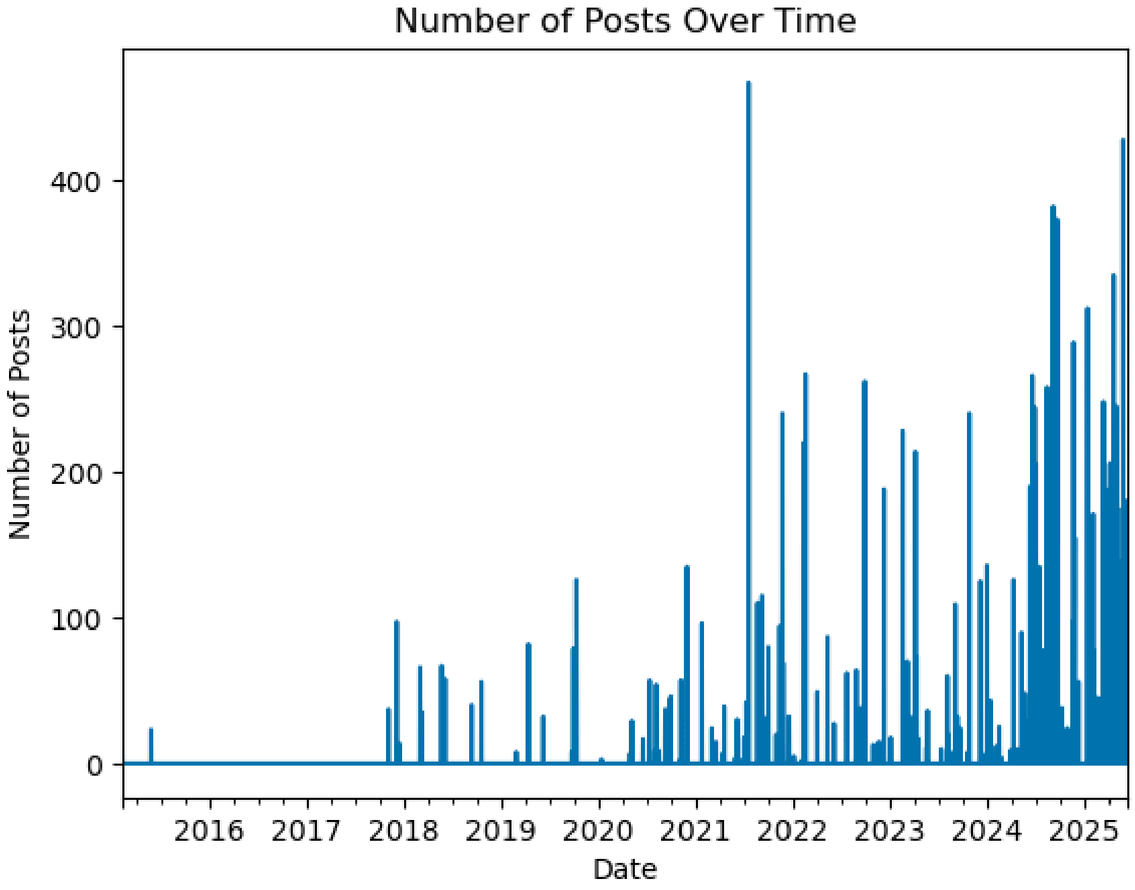
Web Scrapping

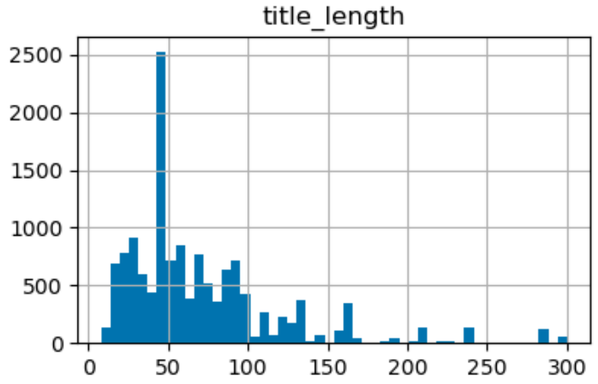
	post_id	title	selftext	post_score	upvote_ratio	num_comments	created_utc	comment_id	comment_body	comment_score	comment_awards	com
0	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0j5twr	Raped. He raped a minor entrusted under his ca...	27	0	
1	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hiscs	Apparently this dude is bro of Dr halina wife ...	74	0	
2	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hs8q8	Like my mom always asked, 'Anak siapa ni?'	18	0	
3	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hpmsh	> She reportedly said the married Amirul Arif ...	28	0	
4	1h704et	Malaysian psychiatrist with 'promising career'...	NaN	137	0.94	46	1.733371e+09	m0hjkkv	nerakazens are doing their job at x. hehehe ...	12	0	
...	...	...	...	...	...	...	...	...	...	...	...	
13718	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😜	NaN	183	0.94	15	1.626157e+09	h51smfn	Hahahaa so funny eh? What about the pakages? T...	-7	0	
13719	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😜	NaN	183	0.94	15	1.626157e+09	h50tvo7	No wonder my package missing!	15	0	
13720	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😜	NaN	183	0.94	15	1.626157e+09	h51h5pi	Haha just allowed it bro	1	0	
13721	oj9ooa	Hidup kena happy.. Rehat jap.. Hilangkan stress 😜	NaN	183	0.94	15	1.626157e+09	h51fswb	Biar lambat asalkan selamat	4	0	
13722	1ar8mxe	Free Stress Management Workshops! 🤩🤯	NaN	8	1.00	1	1.707977e+09	kqoxdwj	UPDATE: We only have very few spaces left for ...	1	0	

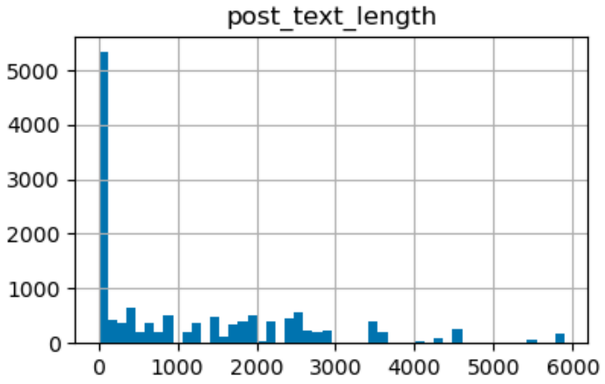
13723 rows × 12 columns

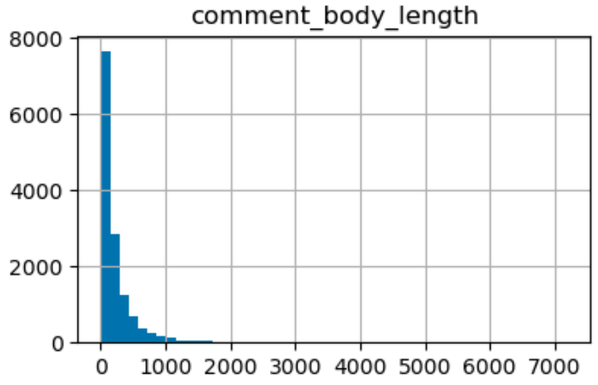
Total data: 13,723 rows

EDA


Number of Posts Over Time
Number of Posts
0
100
200
300
400
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Date

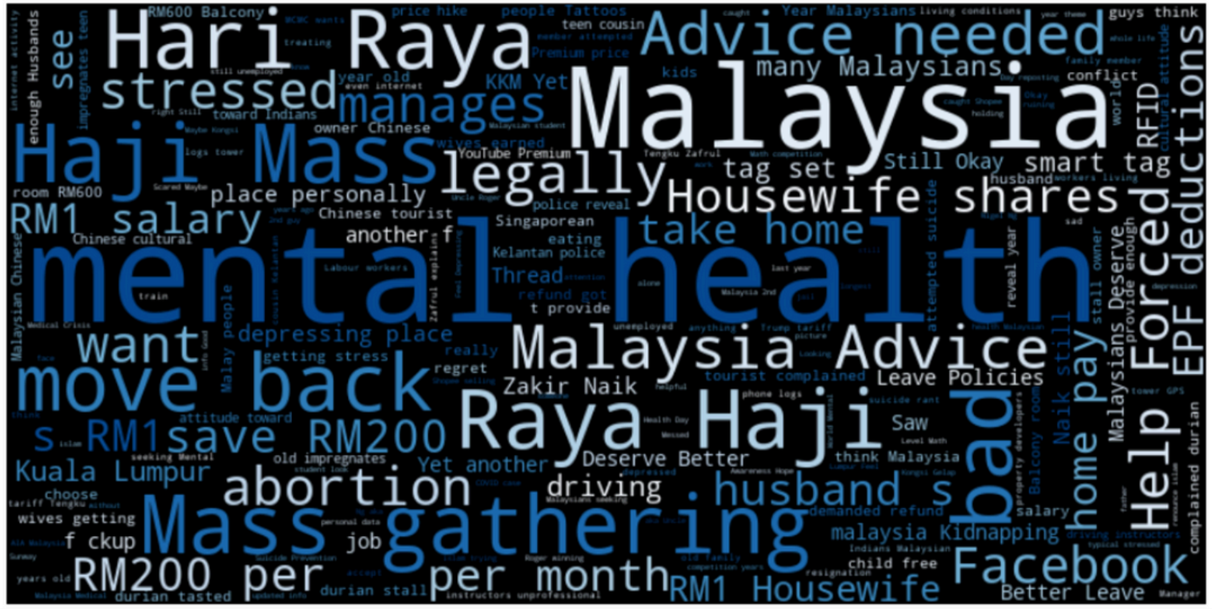

title_length
0
500
1000
1500
2000
2500
0
50
100
150
200
250
300


post_text_length
0
1000
2000
3000
4000
5000
0
1000
2000
3000
4000
5000
6000

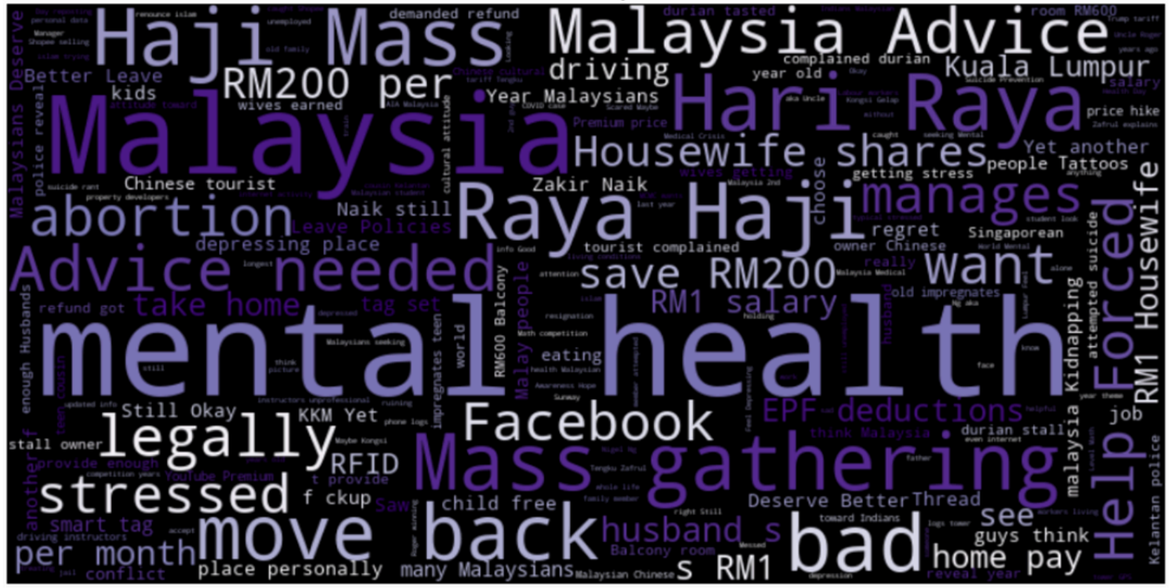

comment_body_length
0
2000
4000
6000
8000
0
1000
2000
3000
4000
5000
6000
7000

Word cloud

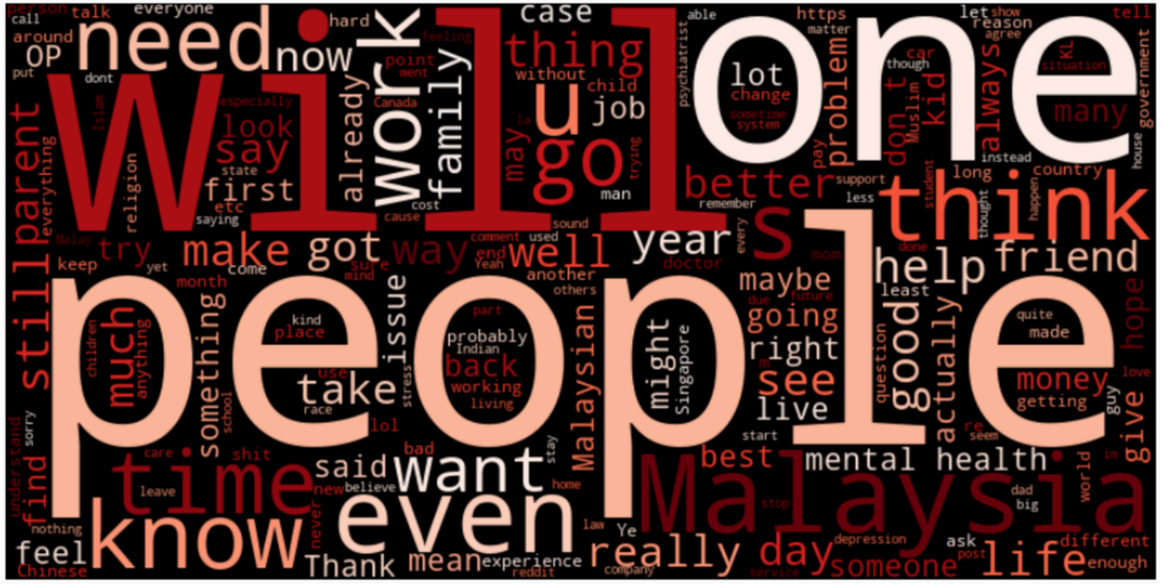
Word Cloud - title



Word Cloud - post text



Word Cloud - comment text



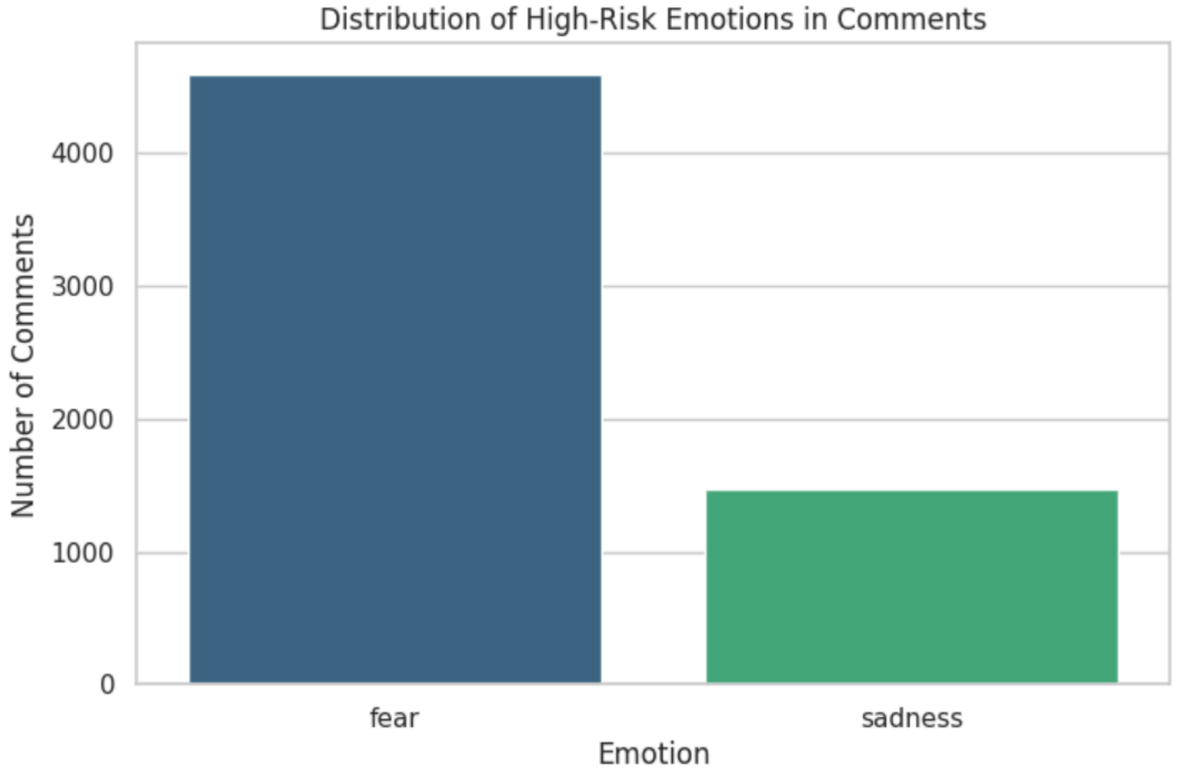
Emotion Classification

	normalized_text	emotion	top_emotion
0	malaysian psychiatrist promising career convic...	{'anger': 0.545, 'love': 0.828, 'sadness': 0.969}	sadness
1	malaysian psychiatrist promising career convic...	{'anger': 0.309, 'love': 0.945, 'sadness': 0.95}	sadness
2	malaysian psychiatrist promising career convic...	{'fear': 0.318, 'love': 0.965, 'sadness': 0.931}	love
3	malaysian psychiatrist promising career convic...	{'fear': 0.956, 'love': 0.591, 'sadness': 0.386}	fear
4	malaysian psychiatrist promising career convic...	{'anger': 0.431, 'love': 0.917, 'sadness': 0.943}	sadness

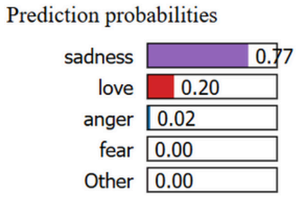
- The model will classify the text with probability of emotion
- Highest probability of emotion will be the top emotion

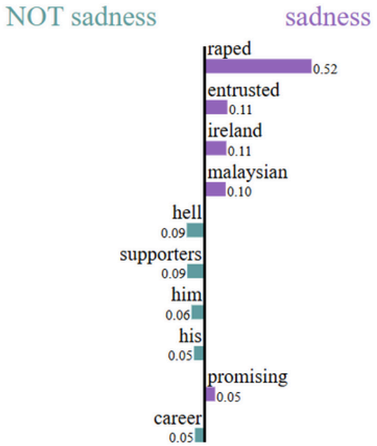
High risk post

6,035 high-risk text



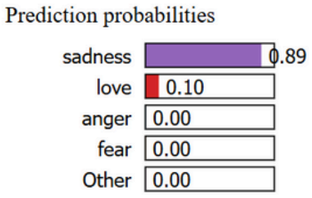
XAI interpretation

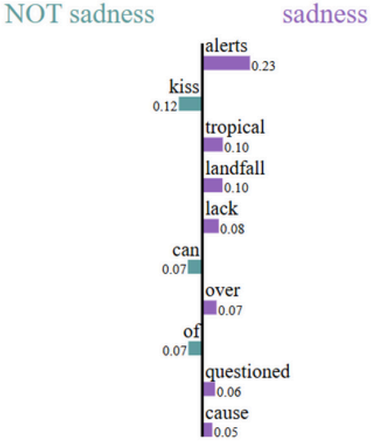




Text with highlighted words

malaysian psychiatrist with promising career convicted of sexually grooming abusing teen girl with mental health issues in ireland raped he raped a minor entrusted under his care rot in hell to him and his supporters





Text with highlighted words

tropical depression s landfall cause of flood disasters govt questioned over lack of alerts the council can kiss my a

Model Evaluation

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.150600	5.962795	0.063179	0.159101
2	0.104500	6.807471	0.059025	0.172983
3	0.068400	7.260326	0.071996	0.173721
4	0.048000	7.244201	0.075202	0.176945
5	0.027400	7.639999	0.066239	0.182285

- Training loss decrease and validation loss increase show the overfitting of the model
- Low accuracy due to domain mismatch and informal language

Conclusion

Conclusion

- The model successfully flagged 6,035 high-risk posts
- LIME helped validate that predictions were based on meaningful cues
- Overfitting was observed during training
- Accuracy is low due to:
  - Lack of true labels
  - Code-switching between Malay/English
  - Informal expressions in Reddit posts

- Future work:

1. manually label a subset of Reddit post
2. Expand data sources other than Reddit like Facebook, Twitter/X

UUNIVERSITI TEKNOLOGI MALAYSIA

univteknologimalaysia utm_my utmofficial

Thank You

www.utm.my

innovative • entrepreneurial • global