

The background is a light gray with a repeating pattern of various tech-related icons. These include gears, arrows, a magnifying glass, a database cylinder labeled 'SQL', a monitor with code, a document with binary code '0110 1010', a puzzle piece, a server rack, a cloud with 'API' and an arrow, and a document with code '</>'.

# **Python for Data Science and AI Presentation II**

**Tan Wei Rong (21224533)**  
**Tan Siao Shuen (20229564)**

# Introduction

# Hugging Face 🤗

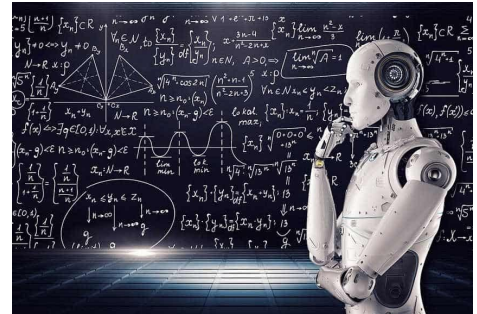
- A free platform for the machine learning community to collaborates on models, datasets, and applications.
- Provides a large collection of pre-trained models for various natural language processing, computer vision, audio, and multimodal tasks.

## The Importance of Hugging Face

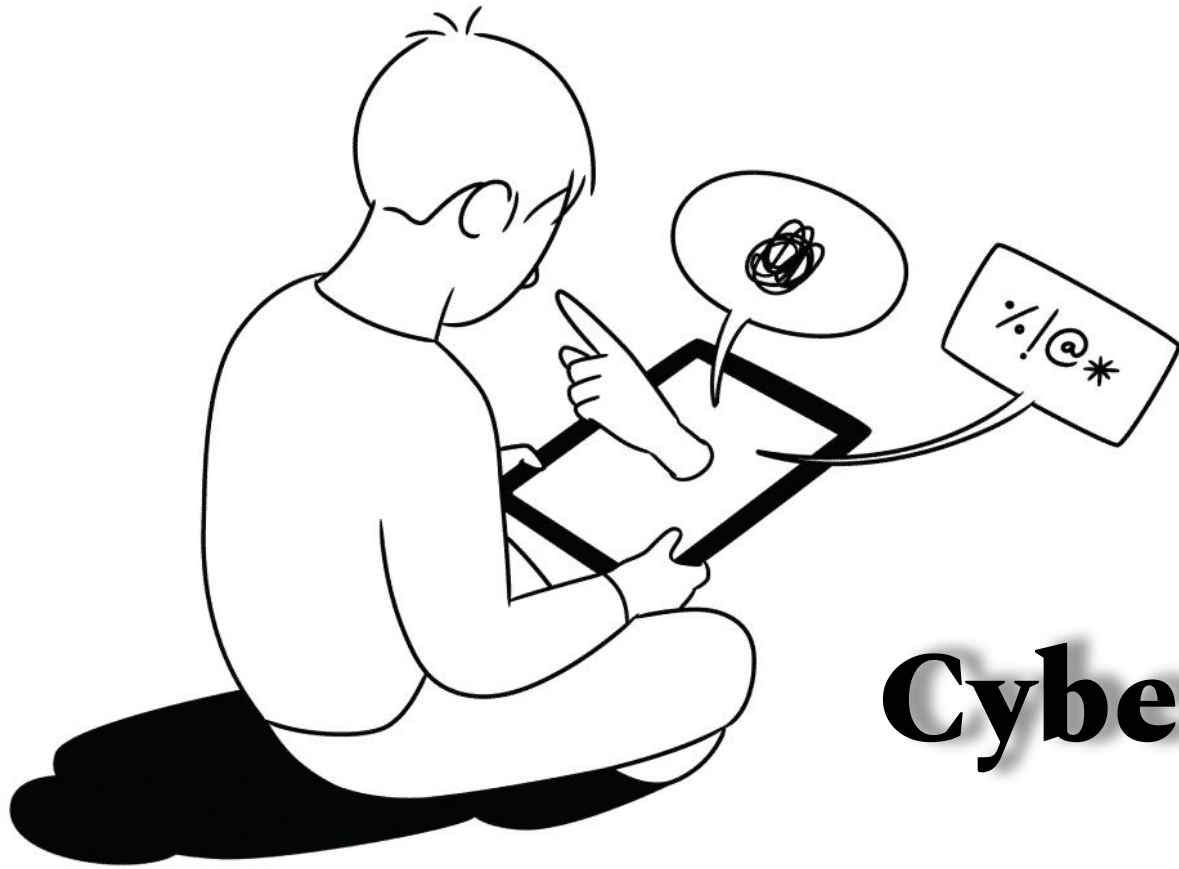
- The community can share, access, and improve machine learning models.
- Reduce the education gaps between professionals and beginners.
- Increase the pace of data science and machine learning related researches as the time of building a model is shortened or even eliminated.

# The Role of Pre-trained Models in DS & AI research field

- Pretrained models could undeniably accelerate data science and AI research.
- Building a model takes a lot of time, thus, being able to access pretrained models created by other people would significantly reduce the time of conducting a research.
- When a fully or partially desired model is done and shared by other people, it would be time-saving as the researchers are not required to build their own model from the scratch.

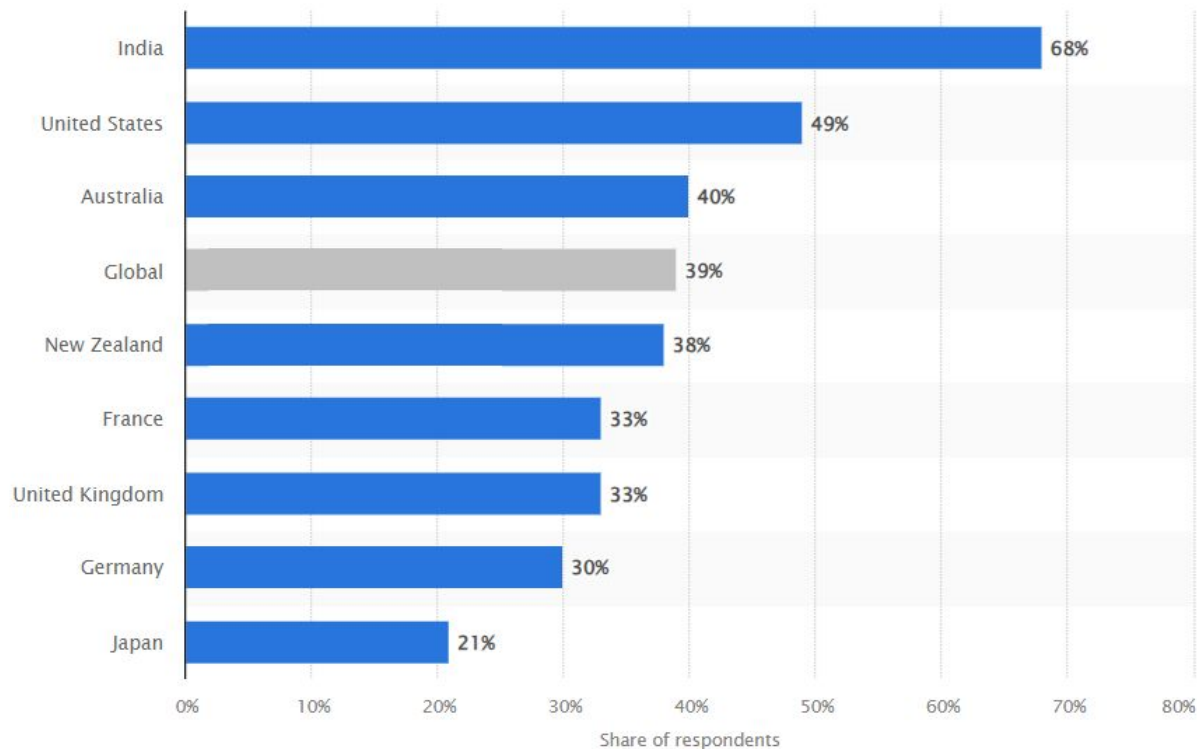


# Problem and Motivation



# Cyberbullying

# Percentage of internet users in selected countries who have ever experienced any cybercrime in 2022 (Petrosyan, 2023)



However...

“Laws against bullying, particularly on cyberbullying, are relatively new and still do not exist everywhere.”

UNICEF (2023)



# The Problem : Cyberbullying

Therefore, we would like to try using AI and machine learning to determine whether a person's comment/post/message is resulting in cyberbullying.

This will be performed by the help of a pretrained model called 'toxic comment model' (by Martin Ha).





# The Importance of Conducting this Research

- To address cyberbullying issue by differentiating toxic/harmful text from the normal ones by using the text classification technique.
- With text being one of the most common types of unstructured data, there are around 80% of all information is unstructured. It is estimated that only 0.5% of unstructured data is being used (Deep Talk, 2021).
- As cyberbullying is a real-world issue, we could make use of the unstructured data in building a text classification model.
- The ideal model after testing and training would not only be able to address cyberbullying but other issues such as hate speech identification and other forms of text classification (such as organizing business surveys and social media accounts) .

# Data and Preprocessing

# Our Dataset

We used a dataset from kaggle comprising of tweets labelled with cyberbullying (1) or not cyberbullying (0) by the author.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import classification_report
%matplotlib inline
```

```
import kaggle
from kaggle.api.kaggle_api_extended import KaggleApi
api = KaggleApi()
api.authenticate()
```

```
x = api.dataset_list(search="cyberbullying tweets")
print(*x)
```

```
soorajtomar/cyberbullying-tweets andrewmvd/cyberbullying-classification yasserhessein/arabic-cyberbullying moneyshot495/siber-z
orbalk syedabbasraza/suspicious-communication-on-social-platforms syedabbasraza/suspicious-tweets munkialbright/classified-twee
ts alanoudaldealij/arabic-cyberbullying-tweets sulimanalmasrey/arabic-tweets-cyberbullying shauryapanpalia/cyberbullying-classi
fication momo12341234/cyberbully-detection-dataset munkialbright/suspicious-tweets haifasaleh/cyberbullying-bystander-dataset-2
023 harsh2345/datacopy noyeemhossain135/cyberbullying-tweets saahir2629/cyberbullying-tweets
```

```
api.dataset_list_files('soorajtomar/cyberbullying-tweets').files
```

```
[CyberBullying Comments Dataset.csv]
```

```
api.dataset_download_file('soorajtomar/cyberbullying-tweets', 'CyberBullying Comments Dataset.csv')
```

```
True
```

# Loading our dataframe

```
df=pd.read_csv('CyberBullying%20Comments%20Dataset.csv')
df
```

	Text	CB_Label
0	damn there is someones nana up here at beach w...	0
1	no kidding! dick clark was a corpse mechanical...	0
2	i read an article on jobros and thought damn w...	0
3	I got one fucking day of sprinkles and now it'...	0
4	I was already listening to Elliott smith and ...	0
...	...	...
11095	"Don't worry you little empty head over it .....	1
11096	"Some of Ya'll are dumb as fuck.... These are ...	1
11097	"Lana, you're so full of shit your eyes are br...	1
11098	"You ain't lying let the @dbeeio61:disqus\xa0\...	1
11099	"Looks like that little Cut-n-paste job has go...	1

11100 rows × 2 columns

```
df.columns
```

```
Index(['Text', 'CB_Label'], dtype='object')
```

```
df['CB_Label'].unique()
```

```
array([0, 1], dtype=int64)
```

```
# Checking for null values
df.isnull().sum()
```

```
Text      0
CB_Label  0
dtype: int64
```

# Preparing our dataframe

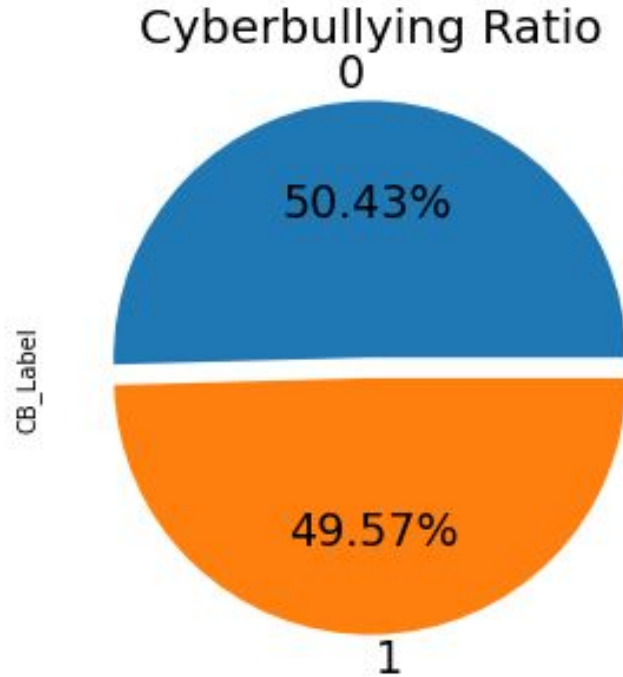
```
for s in range(0,11099):
    text = df['Text'][s]
    length = len(text)
    if length > 511: # getting rid of rows with more than 511 characters
        df = df.drop(s)
df
# 102 rows were removed from the original data
```

	Text	CB_Label
0	damn there is someones nana up here at beach w...	0
1	no kidding! dick clark was a corpse mechanical...	0
2	i read an article on jobros and thought damn w...	0
3	I got one fucking day of sprinkles and now it'...	0
4	I was already listening to Elliott smith and ...	0
...	...	...
11094	"JoeApe - did the room you grow up in have lea...	1
11095	"Don't worry you little empty head over it .....	1
11097	"Lana, you're so full of shit your eyes are br...	1
11098	"You ain't lying let the @dbeio61:disqus\xa0\...	1
11099	"Looks like that little Cut-n-paste job has go...	1

10997 rows × 2 columns

We removed tweets which has more than 512 characters as there as is maximum number of characters which the Hugging Face model can tokenize.

# Final Dataframe



Our final dataframe shows that the amount of tweets that are classified as cyberbullying and non-cyberbullying are approximately half each, which will lessen the risk of producing biased result.

# Model Selection and Training



# Model Selection

## The specific pretrained model ('toxic comment model')

- Fine-tuned
- Accuracy of 94%
- Documentations unavailable on their GitHub repository
- Downloaded at least 1 million times in the past month
- Benefit of doubt on their credibility and reliability.

The model classifies text with a toxicity index, meaning that texts will be classified as either **toxic** or **non-toxic** based on the score predicted by the model.

# Training the model

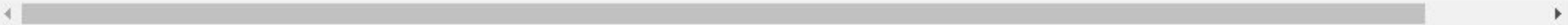
Calling the model through Hugging Face API and the sample output for a random tweet from the dataframe.

```
# Toxic Comments model
import requests

API_URL = "https://api-inference.huggingface.co/models/martin-ha/toxic-comment-model"
headers = {"Authorization": "Bearer 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"}

def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()
```

```
# sample output for a random text from the dataframe
output = query({
    "inputs": 'i read an article on jobros and thought damn we should cash in on our jobro pokemon. Perfect stocking stuffers'
})
output
```



```
[[{'label': 'non-toxic', 'score': 0.7232069969177246},
  {'label': 'toxic', 'score': 0.2767930030822754}]]
```

# Training the model

However, due to the large dataframe and the limits placed by Hugging Face on its API usage, we imported the model for use directly.

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer, TextClassificationPipeline

model_path = "martin-ha/toxic-comment-model"
tokenizer = AutoTokenizer.from_pretrained(model_path)
model = AutoModelForSequenceClassification.from_pretrained(model_path)

pipeline = TextClassificationPipeline(model=model, tokenizer=tokenizer)
print(pipeline('i read an article on jobros and thought damn we should cash in on our jobro pokemon. Perfect stocking stuffers ha
```

# Training the model

```
for i in range(0,10996):
    output = pipeline(df1['Text'][i])
    if output[0]['label'] == 'non-toxic' and output[0]['score'] >= 0.7:
        df1['Cyberbullying'][i] = 0
    elif output[0]['label'] == 'toxic' and output[0]['score'] >= 0.7:
        df1['Cyberbullying'][i] = 1
    else:
        df1['Cyberbullying'][i] = 'Undetermined'
```

df1

	Text	CB_Label	Cyberbullying
0	"Ahhh that would be nice. I'm tired of those c...	1	Undetermined
1	Taking the piss = talking trash perhaps	1	1
2	okay :o	0	0
3	I will vent/blog/whine about my poor character...	0	0
4	"Proof positive that you are idiot"	1	0
...	...	...	...
10992	fuck a sunburn. what are you 8? im on some shr...	0	1
10993	well i hope i will do you guys proud! As soon ...	0	1
10994	oooh i had the sneezes for three days. Dr. sa...	0	0
10995	all these awesome people just flock to me. Gue...	1	1
10996	My kid loves Phantom too. :) She's kind of a m...	0	0

10997 rows × 3 columns

We applied the model for all tweets in the data frame to obtain the predicted outcome.

We also placed a score limit of 0.7, meaning that we only accept tweets with a 'non-toxic' or 'toxic' score of 0.7 and above to prevent ambiguous tweets from affecting our predictions. Tweets that do not meet the limit will be classified as 'Undetermined'.

# Training the model

```
df2 = df1.drop(df1[(df1.Cyberbullying == 'Undetermined')].index)
```

```
df2 # 986 rows with undetermined data were removed
```

	Text	CB_Label	Cyberbullying
1	Taking the piss = talking trash perhaps	1	1
2	okay :o	0	0
3	I will vent/blog/whine about my poor character...	0	0
4	"Proof positive that you are idiot"	1	0
5	"Then go away Nobody here wants to hear your b...	1	1
...	...	...	...
10992	fuck a sunburn. what are you 8? im on some shr...	0	1
10993	well i hope I will do you guys proud! As soon ...	0	1
10994	oohh I had the sneezes for three days. Dr. sa...	0	0
10995	all these awesome people just flock to me. Gue...	1	1
10996	My kid loves Phantom too. :) She's kind of a m...	0	0

10011 rows × 3 columns


	Text	CB_Label	Cyberbullying
1	Taking the piss = talking trash perhaps	yes	yes
2	okay :o	no	no
3	I will vent/blog/whine about my poor character...	no	no
4	"Proof positive that you are idiot"	yes	no
5	"Then go away Nobody here wants to hear your b...	yes	yes
...	...	...	...
10992	fuck a sunburn. what are you 8? im on some shr...	no	yes
10993	well i hope I will do you guys proud! As soon ...	no	yes
10994	oohh I had the sneezes for three days. Dr. sa...	no	no
10995	all these awesome people just flock to me. Gue...	yes	yes
10996	My kid loves Phantom too. :) She's kind of a m...	no	no

10011 rows × 3 columns

CB\_Label : True data  
Cyberbullying : Model predictions

# Model Evaluation

One of the most widely used evaluation matrix for the toxic comment classification is the F1 score (Androcec, 2020).



Evaluation metric	Number of papers
F1 score	15
Accuracy	14
Area Under the ROC Curve (AUC ROC)	9
Custom AUC bias metric	2
Log loss	2
Hamming loss	2
False discovery rate	2
Mean precision	2
Mean recall	2
Pearson correlation coefficients	1
Specificity	1
Mean of the error rates	1
Generalized Mean Bias AUC	1
Subgroup AUC	1
BPSN AUC	1


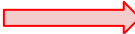


# Model Evaluation using Hugging Face Model

```
# HuggingFace Model results  
print(classification_report(df2['CB_Label'], df2['Cyberbullying'], target_names=['yes', 'no']))
```

	precision	recall	f1-score	support
yes	0.63	0.62	0.62	5052
no	0.62	0.63	0.63	4959
accuracy			0.63	10011
macro avg	0.63	0.63	0.63	10011
weighted avg	0.63	0.63	0.63	10011

# Choosing Our Comparison Algorithm

According to Androcec (2020), these are the algorithms that academic papers have used in toxic comment classification.

	Machine learning method	Number of papers
	Convolutional neural network (CNN)	12
	Logistic regression classifier	9
	Bidirectional long short-term memory (BiLSTM)	8
	Bidirectional Gated Recurrent Unit Networks (Bidirectional GRU)	6
	Long Short Term Memory (LSTM)	5
	Support Vector Machine (SVM)	5
	Bidirectional Encoder Representations from Transformers (BERT)	4
	Naive Bayes	4
	Capsule Network	3
	Random Forest	2
	Decision tree	2
	KNN classification	2
	Gated Recurrent Unit (GRU)	2
	Extreme Gradient Boosting (XGBoost)	2
	Recurrent Neural Network (RNN)	2
	Bi-GRU-LSTM	1
	Gaussian Naive Bayes	1
	Genetic Algorithms (GA)	1
	Partial Classifier Chains (PartCC)	1



# Utilizing other model predictions - KNN

```
# get KNN classification report  
print(classification_report(y_test, KNNpredictions, target_names=['yes', 'no']))
```

	precision	recall	f1-score	support
yes	0.56	0.87	0.68	1652
no	0.70	0.31	0.43	1652
accuracy			0.59	3304
macro avg	0.63	0.59	0.55	3304
weighted avg	0.63	0.59	0.55	3304

# Utilizing other model predictions - SVM

```
# get SVM classification report  
print(classification_report(y_test, SVMpredictions, target_names=['yes', 'no']))
```

	precision	recall	f1-score	support
yes	0.68	0.74	0.71	1652
no	0.71	0.64	0.68	1652
accuracy			0.69	3304
macro avg	0.69	0.69	0.69	3304
weighted avg	0.69	0.69	0.69	3304

# Utilizing other model predictions - Logistic Regression

```
# get logistic regression classification report  
print(classification_report(y_test, LOGpredictions, target_names=['yes', 'no']))
```

	precision	recall	f1-score	support
yes	0.69	0.77	0.73	1652
no	0.74	0.66	0.70	1652
accuracy			0.71	3304
macro avg	0.72	0.71	0.71	3304
weighted avg	0.72	0.71	0.71	3304

# Result Comparison

## Hugging Face

	precision	recall	f1-score	support
yes	0.63	0.62	0.62	5052
no	0.62	0.63	0.63	4959
accuracy			0.63	10011
macro avg	0.63	0.63	0.63	10011
weighted avg	0.63	0.63	0.63	10011

## KNN

	precision	recall	f1-score	support
yes	0.56	0.87	0.68	1652
no	0.70	0.31	0.43	1652
accuracy			0.59	3304
macro avg	0.63	0.59	0.55	3304
weighted avg	0.63	0.59	0.55	3304

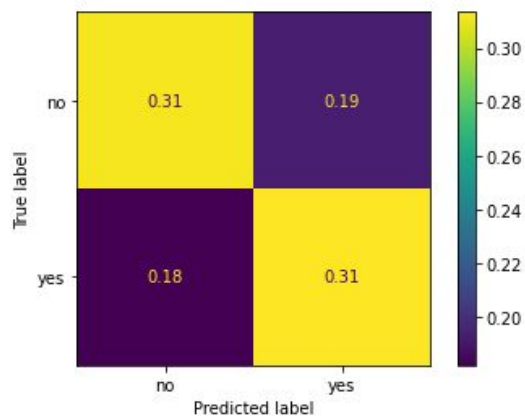
## SVM

	precision	recall	f1-score	support
yes	0.68	0.74	0.71	1652
no	0.71	0.64	0.68	1652
accuracy			0.69	3304
macro avg	0.69	0.69	0.69	3304
weighted avg	0.69	0.69	0.69	3304

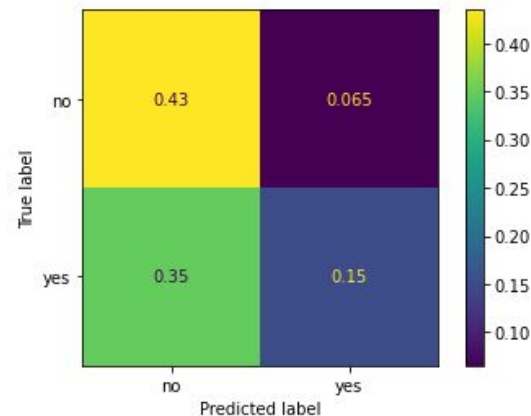
## Logistic Regression

	precision	recall	f1-score	support
yes	0.69	0.77	0.73	1652
no	0.74	0.66	0.70	1652
accuracy			0.71	3304
macro avg	0.72	0.71	0.71	3304
weighted avg	0.72	0.71	0.71	3304

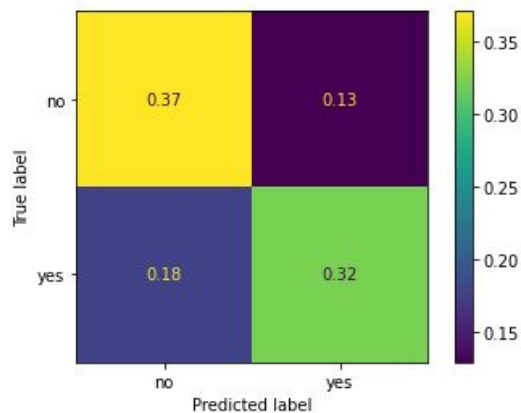
## Hugging Face



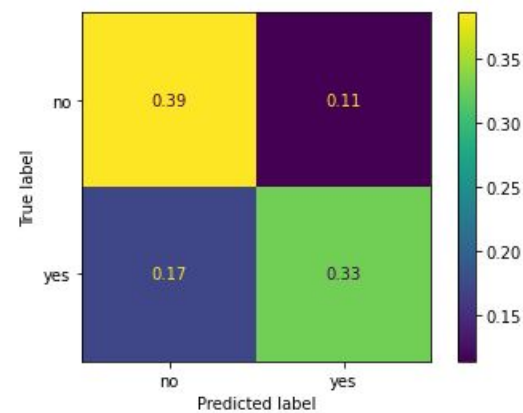
## KNN



## SVM



## Logistic Regression



## Results and Discussion

# Results & Discussion

Model	F1 score
Hugging Face	0.63
KNN	0.59
SVM	0.69
<b>Logistic Regression</b>	<b>0.71</b>

Logistic Regression seems to be a better model for prediction compared to the Hugging Face model.

One of the limitations of the Hugging Face model that is mentioned by the author is that it performs poorly for some comments that mention a specific identity subgroup, such as Muslim.

Another reason for the poor performance might be due to evolution of linguistics as well as the usage of internet slang on social media platforms. The current Hugging Face model might not have enough pre-trained data as well as dictionaries on the use of internet slang to be able to accurately predict whether the tweet is considered to be cyberbullying or not.

# Implications

Propose to be used for user management on social media platforms

- Model helps to distinguish cyberbullies using the platform
- Analyze whether majority of posts from the person are toxic or not
- Restrict usage of the person

Constantly reviewing and updating the prediction model through the online community, allows us to create a more accurate model which we can use to demonstrate the usefulness of AI as a tool in the lawmaking process to legislation.

Enacting relevant laws on cyberbullying will help in maintaining social order as internet users would be more self-conscious while posting, commenting, or messaging using online platforms. However, in order to implement this, it is important to abolish network anonymity so that the identity of cyberbullies are not hidden.

With this, netizens are accountable for their actions online and practise caution while using the internet.



## Conclusion and Future Work

# Main Takeaways

- Current Natural Language Processing models might not work well with social media platforms
- Understand the pre-trained model and their limitations
- Investigate which data the pre-trained model had used for training

In our case, we hypothesized that the Hugging Face did not have training data which consists of internet slang.

Hence the pre-trained model is unable to accurately predict whether the comment is considered to be cyberbullying or not.

# Potential Future Directions

In order to strengthen Natural Language Processing models,

- Develop a dictionary for internet slang
- Constantly update the dictionary

so that more models can utilize the dictionary for more accurate results and predictions.

Additionally, we should not blindly trust pre-trained models available on Hugging Face to be entirely accurate. It is important to be able to ascertain the strengths and weaknesses of the models.

When working with pre-trained models, we should cross check our own models with the pre-trained models from Hugging Face and compare their predictions and accuracies. The ultimate goal is to figure out the optimal model which best fits a specific dataset.

Thank You!

Thank you



# References

- Androcec, D. (2020). Machine learning methods for toxic comment classification: A systematic review. *Acta Universitatis Sapientiae Informatica*, 12 (2), 205-216. <https://doi.org/10.2478/ausi-2020-0012>
- Deep Talk. (2021). 80% of the world's data is unstructured. Medium. <https://deep-talk.medium.com/80-of-the-worlds-data-is-unstructured-7278e2ba6b73>
- Tomar, S. (2023). *Cyberbullying Tweets*. Kaggle. <https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets>
- Ha, M. (2021). *Toxic Comment Model*. Hugging Face. <https://huggingface.co/martin-ha/toxic-comment-model>
- Hugging Face. <https://huggingface.co/>
- Petrosyan, A. (2023). *Cybercrime encounter rate in selected countries 2022*. Statista. <https://www.statista.com/statistics/194133/cybercrime-rate-in-selected-countries/#statisticContainer>
- UNICEF (2023, February). *Cyberbullying: What is it and how to stop it*. United Nations International Children's Emergency Fund. <https://www.unicef.org/end-violence/how-to-stop-cyberbullying>