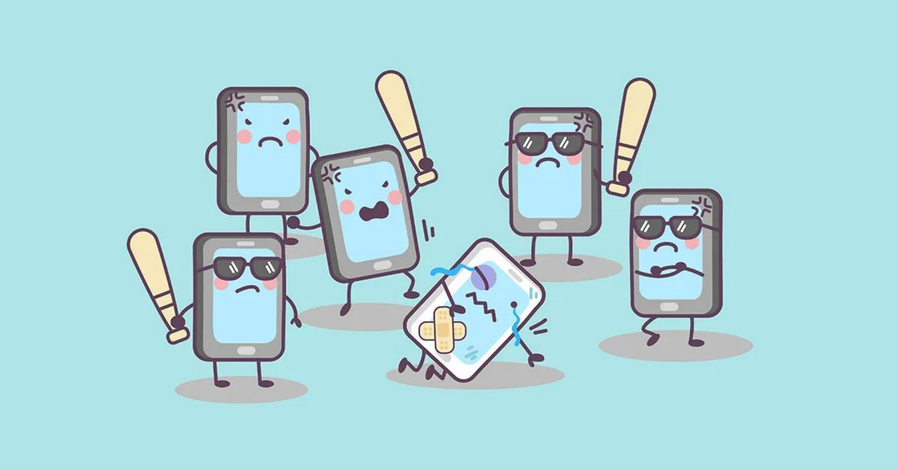
General Assembly Singapore, 2025

Seth Lee

DABPT01 Capstone Assignment

Final Presentation

THE PROBLEM



Background

The online space can be a dangerous place to be, especially since anyone and everyone can partake in it

Toxic online comments are a result of online anonymity, and can be dangerous for unsuitable audiences if left unchecked for too long

Project Objective

By way of a natural language processing (NLP) model, ascertain if any given online comment, could be classified as a 'problematic' / 'negative' one, or left alone without consequence

Applications

Moderation of any website / application with:

- Free-text comment section
- Message boards
- Chatbots

6 FLAGS

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

159,571

Total no. of social media comments analysed

10.16%

Proportion of comments with **≥1** toxicity flags

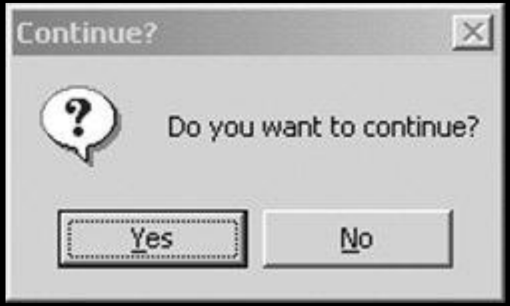
DISCLAIMER!!!

Presentation material beyond this point, while constructed strictly for academic purposes, may be considered <u>**offensive, controversial, or triggering**</u> to some viewers.

This includes, but is not limited to, themes of:

- **Violence**
- **Discrimination**
- **Sexually suggestive / explicit content**
- **Strong language**

<u>**Viewer discretion is advised.**</u> If you are sensitive to such topics, you may wish to consider this before proceeding.



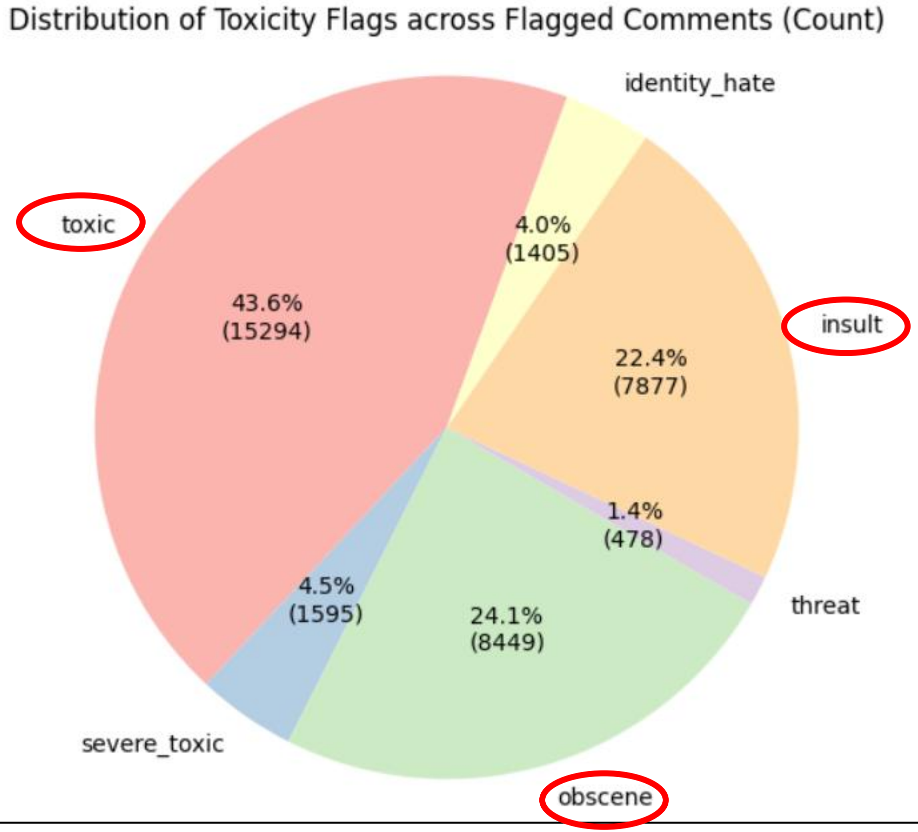
toxic

severe_toxic

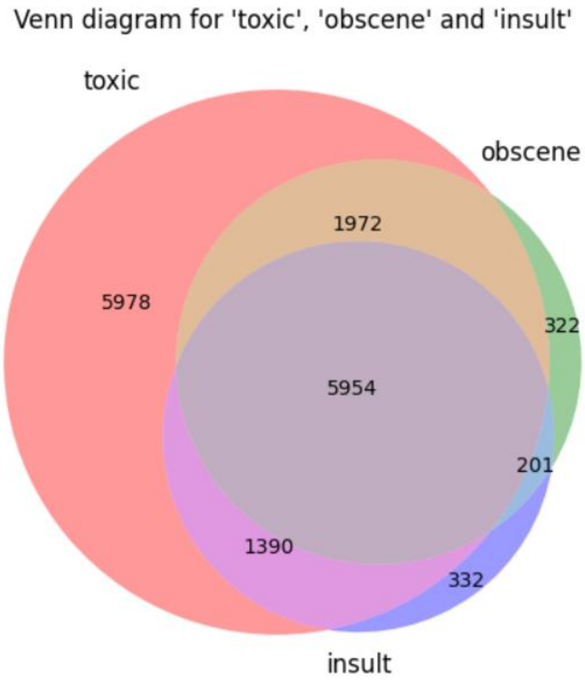
obscene

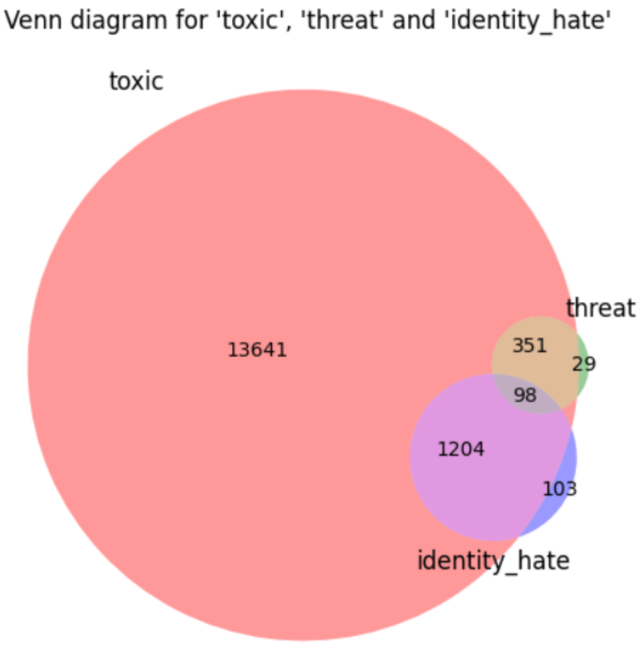
insult

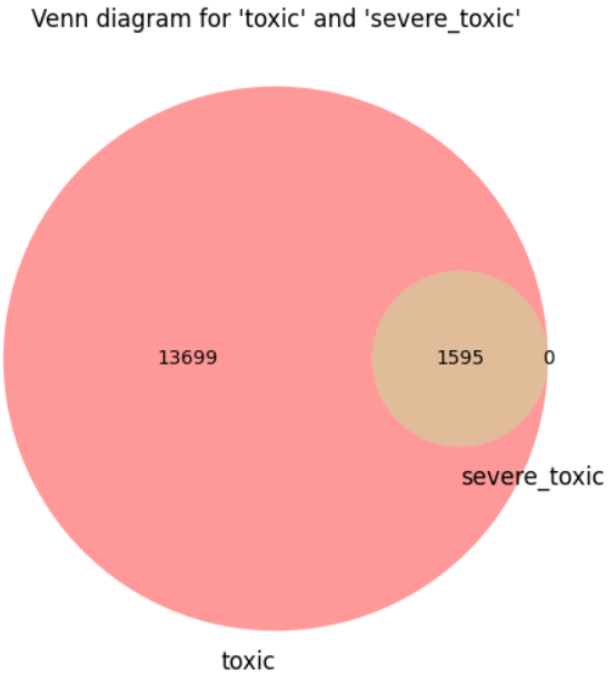
EXPLORATORY DATA ANALYSIS (EDA)



HIGH DEGREE OF OVERLAP BETWEEN FLAGS?







Urgency of corrective action?

Value of multi-label data?

- In real-world context, swift corrective action on problematic social media comments would be of higher importance
- Toxicity labels can be condensed into one single flag (eg. ‘flagged’) and analysed on binary basis
  - Problematic (‘flagged’ = 1)
  - Innocent (‘flagged’ = 0)

5,461,520

Total no. of tokens generated from 159,571 comments

TRUE +VE

"i am going to pee on you"

TRUE -VE

"yep it was cut around _ seconds however it has been released uncut now"

FALSE +VE

"burial where did he die where is he buried"

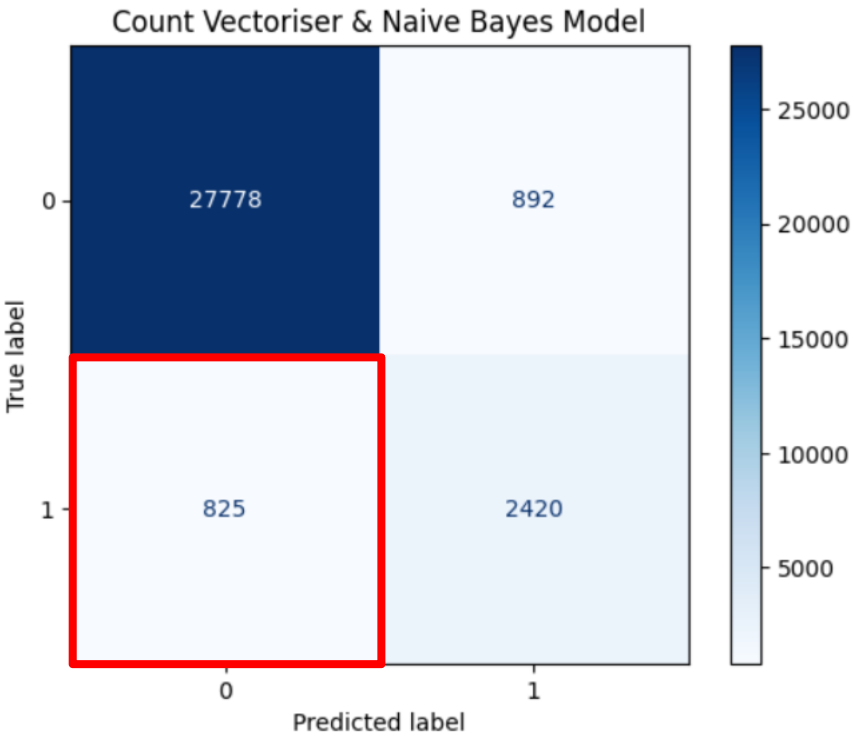
"would care less good luck have fun enjoy your life"

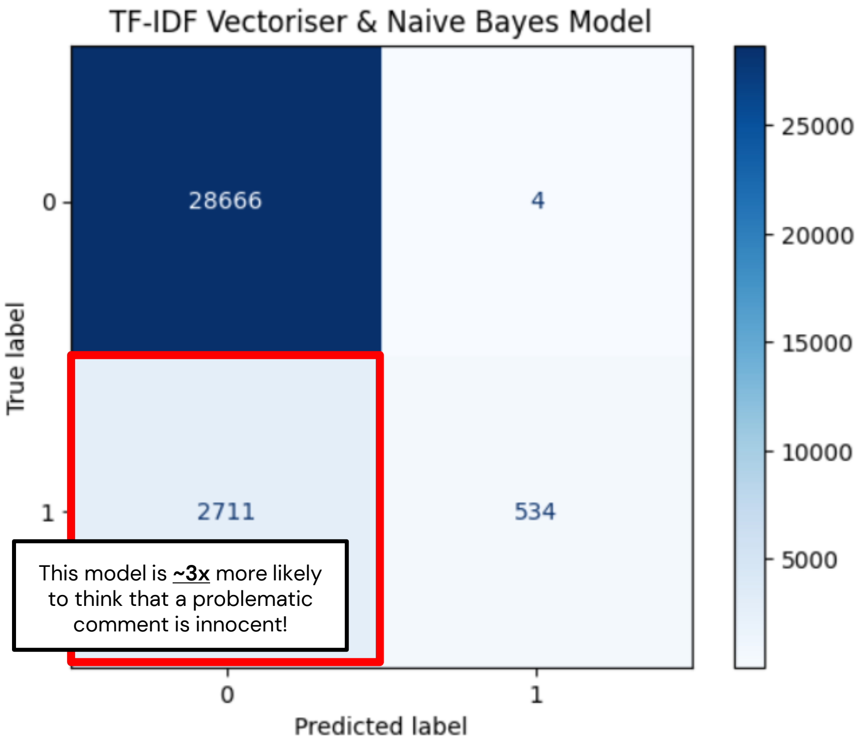
FALSE -VE

"i am having my period"

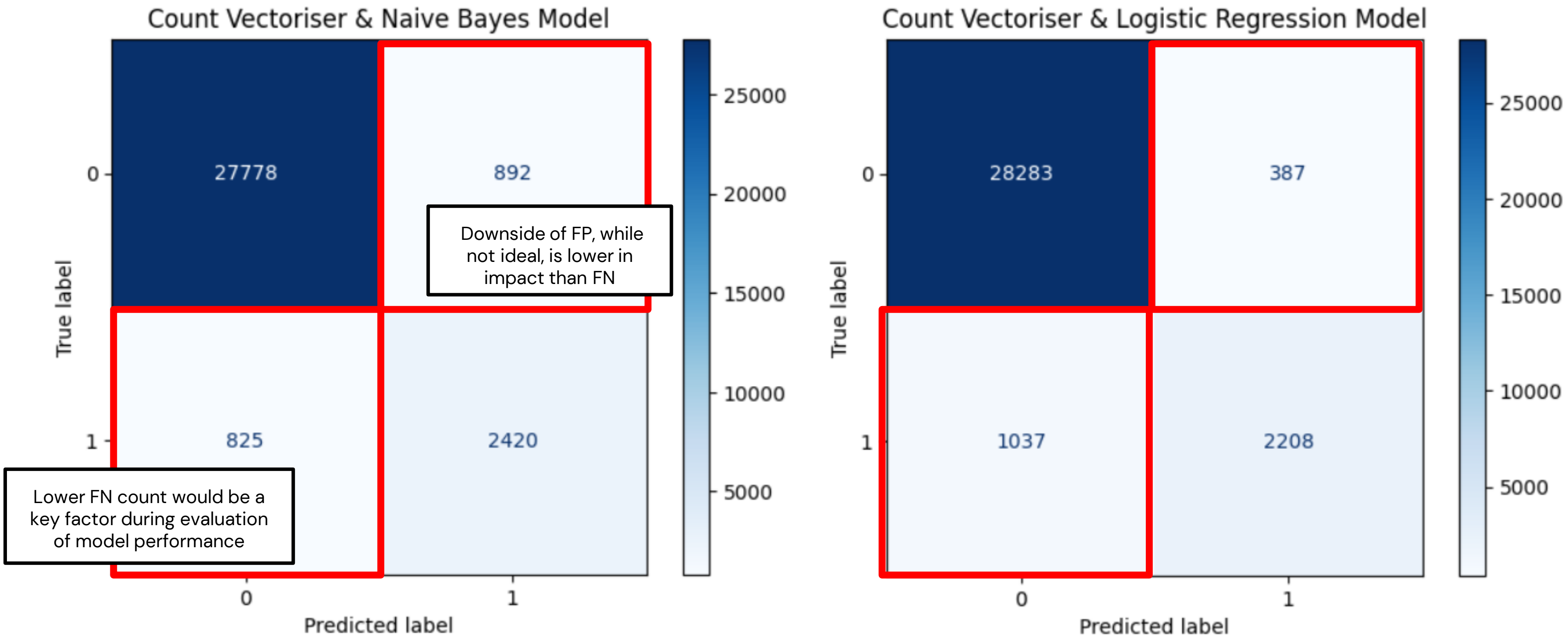
"do it and i will cut you"

WHY IS FALSE -VE SIGNIFICANT HERE?





BEST PERFORMING VECTOR-MODEL COMBOS

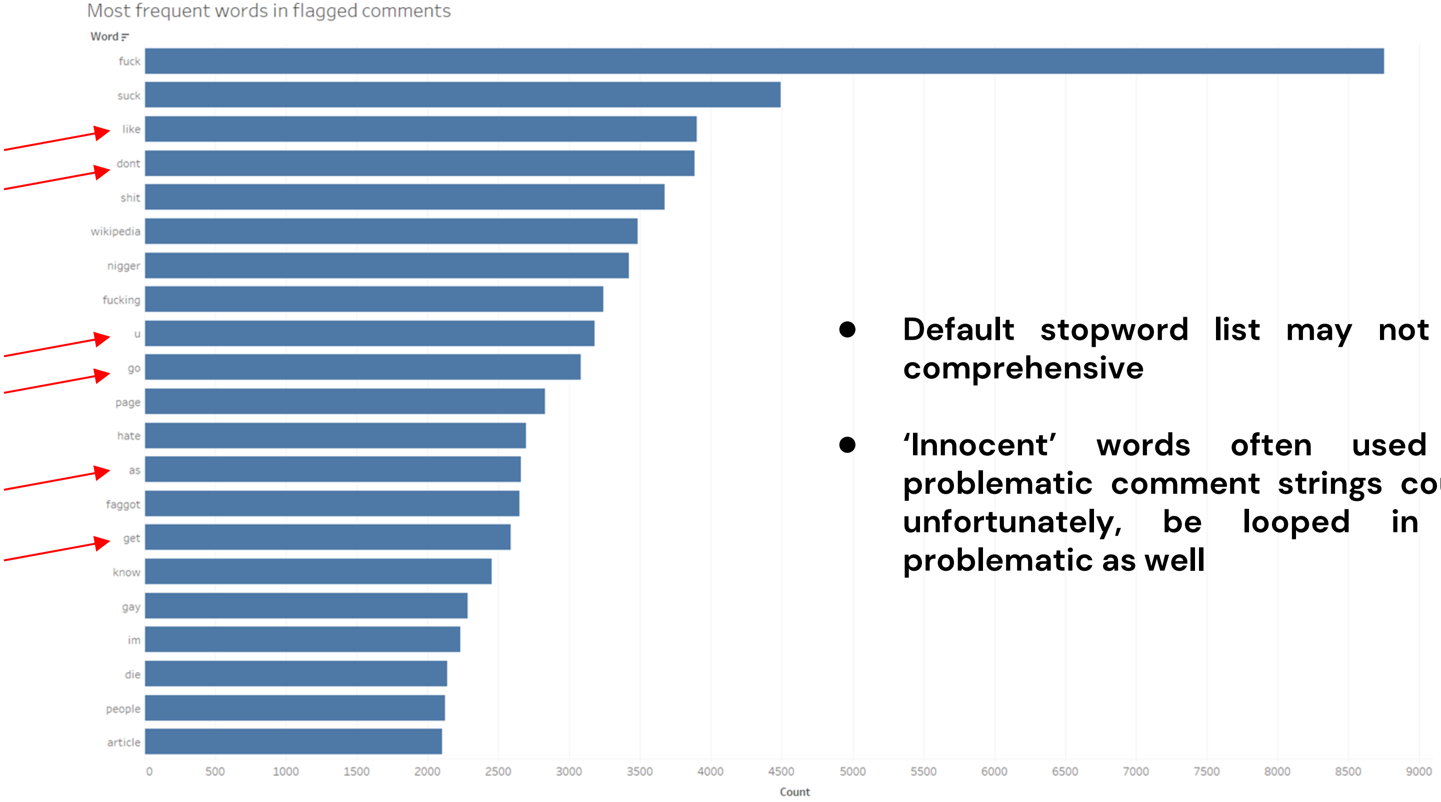


VECTORISER-MODEL PERFORMANCE

Vectoriser	Model	Accuracy	Precision	Recall	F1-Score
Count	Dummy (Baseline)	82.07%	11.60%	11.53%	11.56%
Count	Logistic Regression	95.54%	85.09%	68.04%	**75.62%**
Count	Naive Bayes	94.62%	73.07%	**74.58%**	73.81%
TF-IDF	Dummy (Baseline)	81.54%	9.77%	9.89%	9.83%
TF-IDF	Logistic Regression	**95.60%**	92.01%	62.10%	74.15%
TF-IDF	Naive Bayes	91.49%	**99.26%**	16.46%	28.23%

Metrics above are measured against the vectoriser-model correctly labelling a (‘flagged’ = 1) comment

ANALYSIS LIMITATIONS



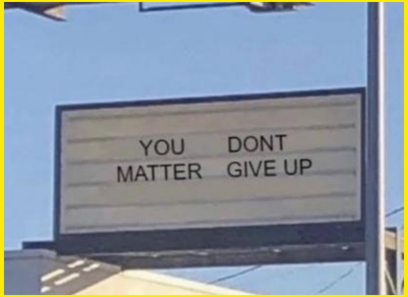
- Default stopword list may not be comprehensive
- ‘Innocent’ words often used in problematic comment strings could, unfortunately, be looped in as problematic as well

ANALYSIS LIMITATIONS

Absence of Context

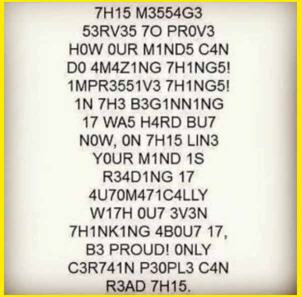
Each word is analysed as is

Unigram Analysis



Word order is ignored

Creative Spelling



Social media users, while trying to circumvent rules and regulations, could still spell offensive words with similar alphanumeric characters

Boundless Nature of Language

If target word was not fed into machine learning model during training, prediction will fail during test phase

Some areas to look into as part of future work...

- Customised / updated stopword library for text cleaning
- Look into more sophisticated models involving multi-label classification, where more information can be gleaned from predictions (eg. OneVsRestClassifier)
- Look into *n*-gram analysis (bi-gram, tri-gram)

Digitalisation breeds anonymity.
Anonymity emboldens recklessness.
Recklessness begets suffering.

Be kind, even when faceless.

Thank You