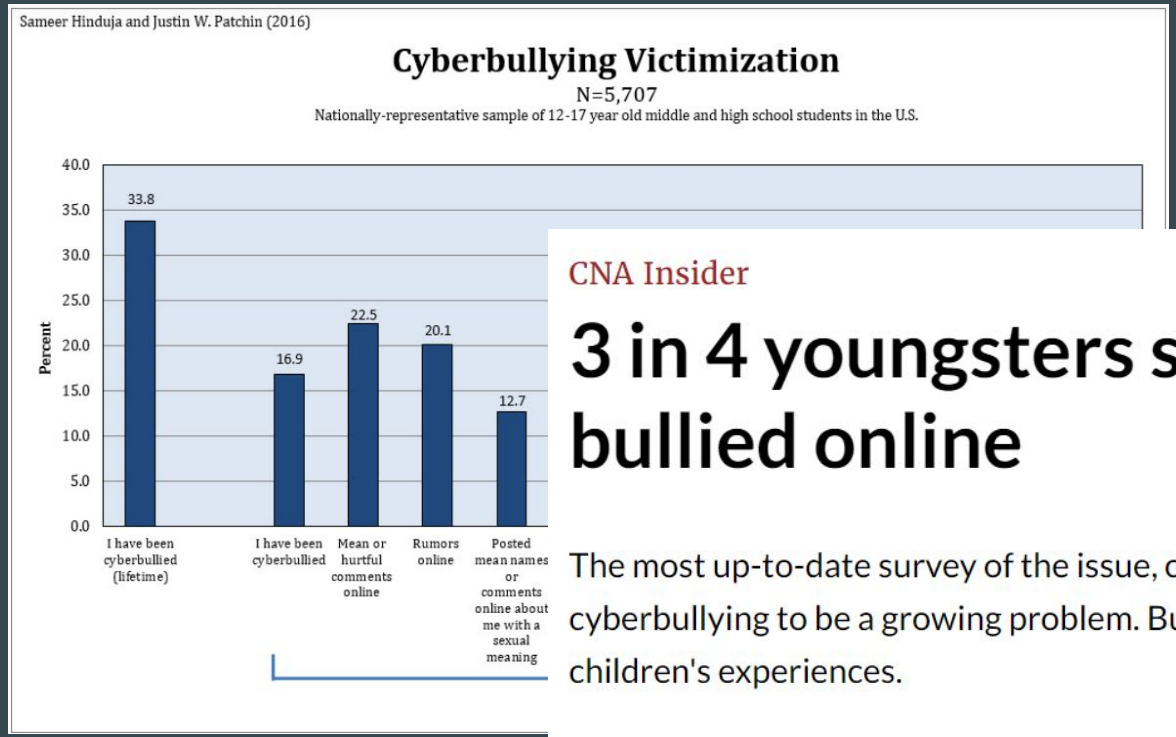
Reduce Cyberbullying via Curbing of Toxic Comments

Nicholas Lim

Content

- Background & Motivation
- Problem Statement
- About Dataset
- Data Cleaning
- Visualizations
- Model Approach
- Evaluation of Models
- Conclusion

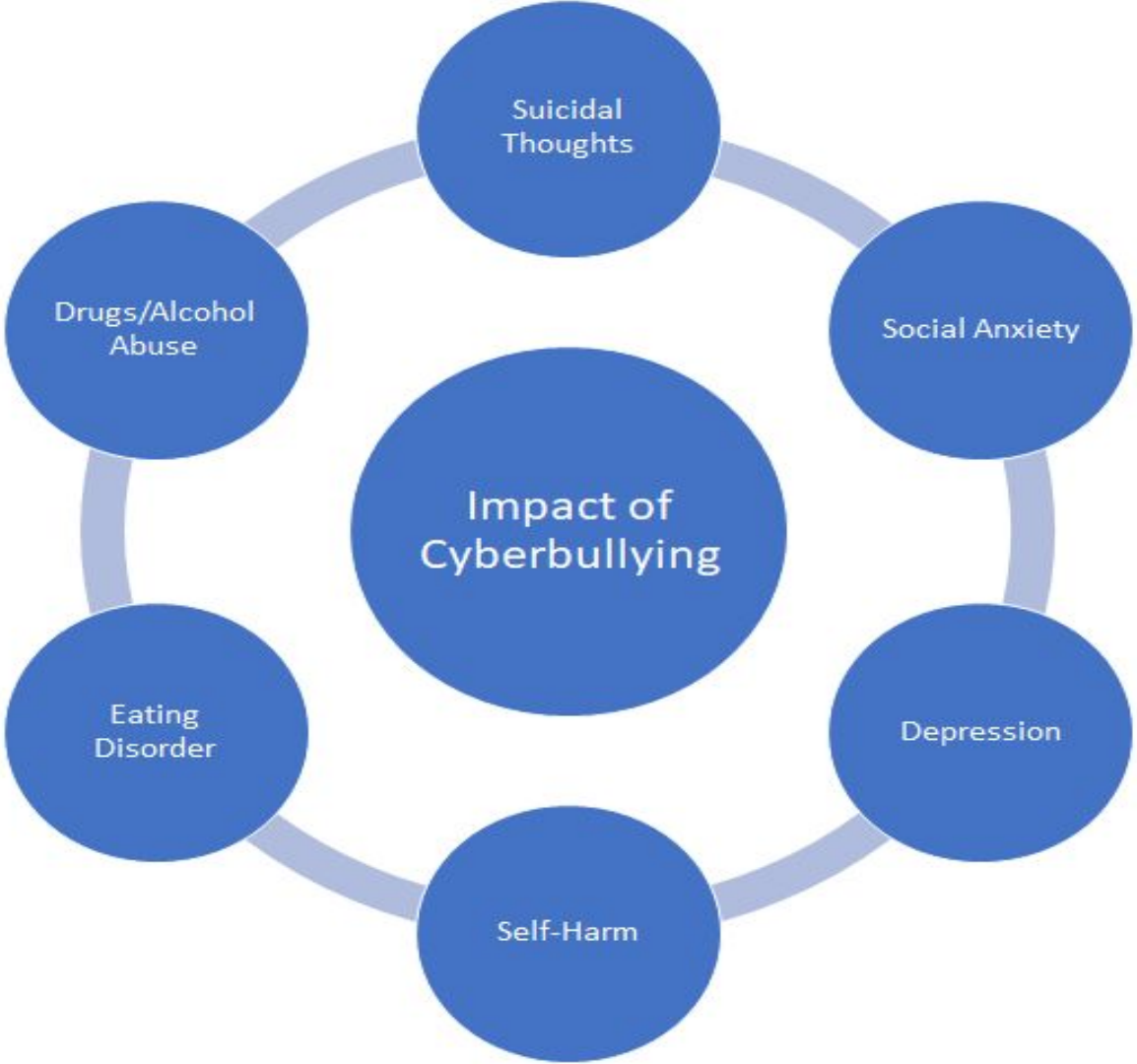
Cyberbullying - a global issue



CNA Insider

3 in 4 youngsters say they have been bullied online

The most up-to-date survey of the issue, commissioned by Talking Point, finds cyberbullying to be a growing problem. But parents may be none the wiser about their children's experiences.


Suicidal Thoughts
Social Anxiety
Depression
Self-Harm
Eating Disorder
Drugs/Alcohol Abuse
Impact of Cyberbullying

Problem Statement

Given an accurately labelled dataset of toxic comments, the task is to determine if a given comment is considered offensive to another person.

This can also be further extended to determine the target of the offensive comment.

About the dataset

- ~1.8m comments with 45 feature columns
- Comments are given 2 types of label:
  - Toxicity
  - Identity Mentions
- Taken from the following Kaggle competition: Jigsaw Unintended Bias in Toxicity Classification

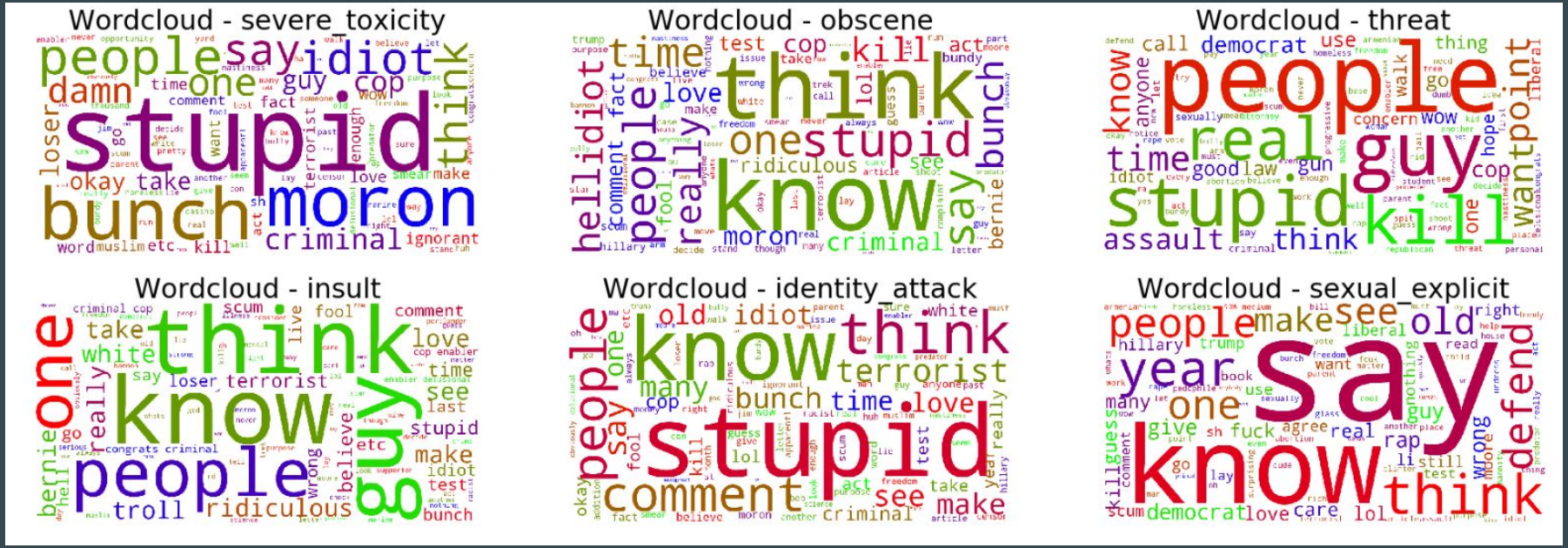
Data Cleaning

Strategy:

- Labels with Null values will be filled with 0
- Lemmatization of words using TextBlob Lemmatization library
- Removal of stop words using NLTK's stopword corpus
- Duplicated comments with same labels will be dropped

A visualization of toxic comments...

Visualizations across the different labels...



Model Approach

2 separate models will be done to tackle the problem

- A binary classification to first determine whether a comment is offensive
- A multi-label classification to determine the type of offensive comments

Model Evaluation

For Binary Classification: (Score based on recall)

	Precision	Recall	f1-Score
Non-Toxic	0.85	0.72	0.78
Toxic	0.72	0.85	0.78

Multinomial NB with CountVectorizer

	Precision	Recall	f1-Score
Non-Toxic	0.84	0.90	0.87
Toxic	0.87	0.80	0.83

Logistic Regression with TFIDF -Vectorizer

	Precision	Recall	f1-Score
Non-Toxic	0.84	0.90	0.87
Toxic	0.87	0.79	0.83

Logistic Regression with CountVectorizer

For Multi-Label Classification:

	Precision	Recall	f1-Score
Obscene	0.68	0.49	0.57
Insult	0.86	0.90	0.88
Identity Attack	0.71	0.48	0.57
Others	0.58	0.33	0.42

OneVsRest with CountVectorizer

	Precision	Recall	f1-Score
Obscene	0.76	0.36	0.49
Insult	0.85	0.93	0.89
Identity Attack	0.75	0.49	0.60
Others	0.72	0.28	0.40

OneVsRest with TFIDF Vectorizer

	Precision	Recall	f1-Score
Obscene	0.67	0.49	0.56
Insult	0.86	0.90	0.88
Identity Attack	0.70	0.49	0.58
Others	0.57	0.38	0.46

Classifier Class with CountVectorizer

	Precision	Recall	f1-Score
Obscene	0.75	0.35	0.47
Insult	0.85	0.93	0.89
Identity Attack	0.75	0.52	0.61
Others	0.70	0.36	0.47

Classifier Class with TFIDF Vectorizer

Conclusion & Recommendations

- Binary classification of toxic comments work well when we under-sample the non-toxic comments and using Logistic Regression with TDIDF-Vectorizer
- Multi-label classification was able to classify insult comments accurately but having rather mixed results when assigning the remaining labels

Limitations of project:

- Hardware of machine used
- Time factor (6 weeks)
- Budget limitation

We can further expand the scope of the project to include the following:

- Applying deep learning to perform unsupervised learning
- Gather more data on toxic comments with more related labels
- Expand the multi-label classification to identity labels

THANK YOU!

...