

---

Princewill Okosun  
&  
Dr Jeffrey Ray

---

# TOWARDS A CLEAN SOCIAL ENVIRONMENT USING MACHINE LEARNING

---



STUDENT NAME:  
STUDENT NUMBER:  
COURSE NAME:  
DEPARTMENT:  
COURSE CODE:  
SUPERVISOR:  
DATE OF SUBMISSION:

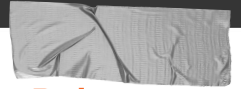
Princewill Okosun Ehimare  
201569621  
Master's in Advanced Computer Science  
Department of Computer Science  
COMP 702  
Dr. Jeffrey Ray  
08-07-2022



—

According to a Pew Research Centre survey [1] four in ten internet users have experienced online harassment with

**45% of these people experiencing heightened levels of hate that could peradventure lead to mental breakdown [1].**



### Main Point

The internet has championed human connectedness but it is being overridden by prevailing **toxicity**.



# Introduction

**The internet** has brought about connectedness and has been a key element in the advancement of technology, however, it is insidiously getting noxious.

→ **Problem Identification**

Toxic culture.

→ **Current Procedures**

Content moderation and regulations.

→ **Problem Resolution**

Artificial Intelligent systems.

—

The internet was designed to be egalitarian in nature, yet, 41% of women who have experienced the brunt of the internet's toxicity say that their encounter had them feeling their physical safety was threatened [2].



#### Main Point

The internet is becoming unbecoming so much so users experience encounters where they are left worried for their **safety**.



# Background Statement

## **Ambiguity in regulations and anonymity of users**

are issues that make regulating internet comments, cyberbullying and internet hate a difficult problem to address.

### → **What is the state of the web?**

Highlight the unusual levels of toxicity on social media platforms.

### → **Why might this be the case?**

Ambiguity, anonymity and inadequate moderation.

### → **The role of Machine Learning**

Automate moderation in order to preserve the mental state of human moderators.

## Objective of the project

The main objective of this project is to raise awareness for the prevailing levels of toxicity in present day internet spaces and social media primarily.





# Specific Aims

Specific aims of this project include but are not limited to:

- **Raise awareness of the volume of hate speech on the internet.**
- **Highlight the consequential impact of internet hate and online abuse.**
- **Bring to light the shortcomings of human moderators and their mental health battles.**
- **Encourage research in sentiment analysis for internet moderation.**
- **Build a word2vec or BERT model for feature extraction.**
- **Identify and build intelligent computational models to categorize hate speech.**
- **Compare model performance using evaluative metrics and against SOTA models.**



As this project categorises hate speech, the data used is limited to textual data. The

## **ANALYTICS**

**VIDYHA** 32k toxic twitter comment dataset is used in the training of our models [3].



### **Main Point**

Models are trained on toxic comments dataset.



# Preprocessing

This preprocessing would involve cleaning the dataset in order to retain salient features that would assist in training the model to predict accurately between hate speech and regular tweets.

## → NLTK

We utilise python's Natural Language Toolkit (NLTK) library in this preprocessing stage of our data. NLTK is a powerful library used in the pre-processing of data for NLP tasks. It also comes with an inbuilt sentiment analyser which outputs the polarity and subjectivity of textual data [4].



To visualise these insights, wordclouds [5] would be used in creating visual representations of relevant words used in hate speech and normal speech.

# Word Embeddings.


This project would use either Word2vec or GloVe for feature extraction.

Features are extracted from the textual data in numerical representations as the machine learning algorithms to be used learn patterns from numeric data [6][7].

## Comparative analysis

The results are compared against their F-scores or using a ROC graph which performs as an excellent evaluative technique in binary classification tasks [8].



A hand holding a smartphone against a blurred red background. The text is overlaid on the left side of the image.

The noxious nature of the dataset used in this project demands elevated levels of protection to prevent unwarranted access to the data as there could be damaging consequences to a person's present mental state.



#### Main Point

Data is abstracted away to prevent unwanted consequences for all external parties.



# Ethics

In addressing a societal issue strongly related to ethics and morality, this project aims to meet the highest computer science ethical standards.

As such the following are carefully considered:

→ **The dataset**

Demands elevated levels of protection.

→ **Impacted Persons**

Since this research could impact the lives of external entities it aims to be clear and honest through providing figures and facts even if they are of a grim nature.

# Project Plan

## June Milestone

Background Reading and Literature Review.

June-July

## August Milestone

Software implementation, Testing, and Presentation

August-September

## July Milestone

Oral Presentation.

## September Milestone

Write-up of dissertation



# Risks and Contingency

To ensure successful completion of this project it is paramount that all risks are identified and assessed. Some risks identified are:

- **Ailing physical or mental health.**
- **Requirement of high-cost pay-to-use online resources.**
- **Incorrect software implementation schedule.**
- **Imbalance in dataset leading to inferior performance.**





## REFERENCES

[1] M. DUGGAN, "Pew Research Center Online Harassment," Pew Research Center, Washington, DC, 2014

[2] A. Dhrodia, "Unsocial media: A toxic place for women," IPPR Progressive Review , 2018.

[3]  
<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>, "Twitter Sentiment Analysis," Analytics Vidhya, 2018.

[4] NLTK Project, "NLTK," NLTK Project, 2022. [Online]. Available: <https://www.nltk.org/api/nltk.html>. [Accessed 3 July 2022].

[5] A. Mueller, "WordCloud for Python documentation," 2020.

[6] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in," California, 2013.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in," California, 2013.

[8] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.," Journal of Pattern Recognition Letters – Special issue in Roc analysis in pattern recognition archive, 2006.