Dr Jeffrey Ray
&
Dr Alexei Lisitsa
&
Princewill Okosun

TOWARDS A CLEAN SOCIAL ENVIRONMENT USING MACHINE LEARNING

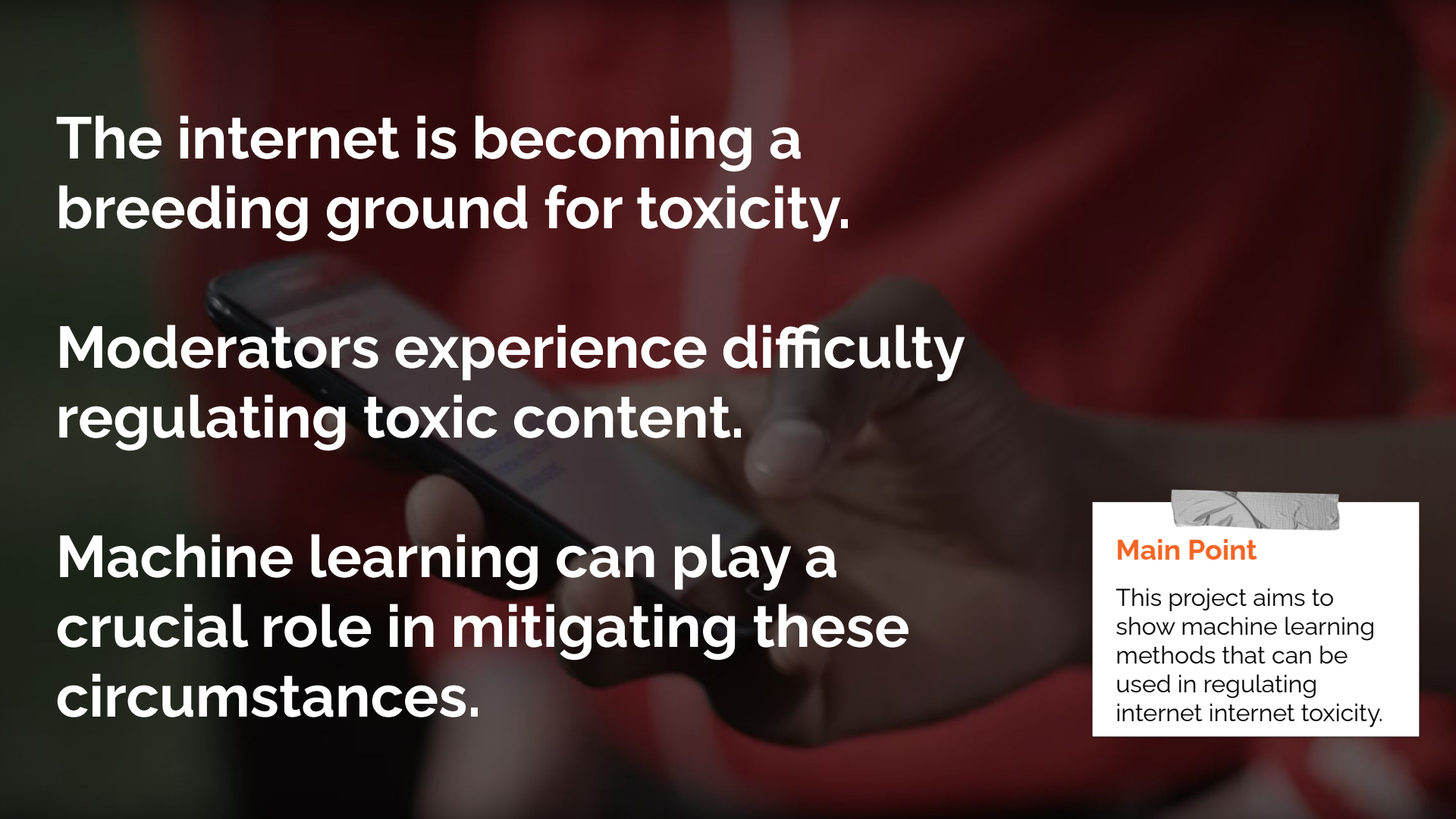
STUDENT NAME: Princewill Okosun Ehimare
STUDENT NUMBER: 201569621
COURSE NAME: Master's in Advanced Computer Science
DEPARTMENT: Department of Computer Science
COURSE CODE: COMP 702
SUPERVISORS: Dr. Jeffrey Ray & Dr. Alexei Lisitsa
DATE OF SUBMISSION: 26-08-2022

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

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

Main Point

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



The internet is becoming a breeding ground for toxicity.

Moderators experience difficulty regulating toxic content.

Machine learning can play a crucial role in mitigating these circumstances.

Main Point

This project aims to show machine learning methods that can be used in regulating internet internet toxicity.

TOWARDS A CLEAN SOCIAL ENVIRONMENT

To **“Research and Implement Technologies”** aimed at addressing the toxic nature of internet comments. This project achieves:

➔ **Research**

Provides data analysis regarding toxic internet culture.

➔ **Data Analysis & Visualization**

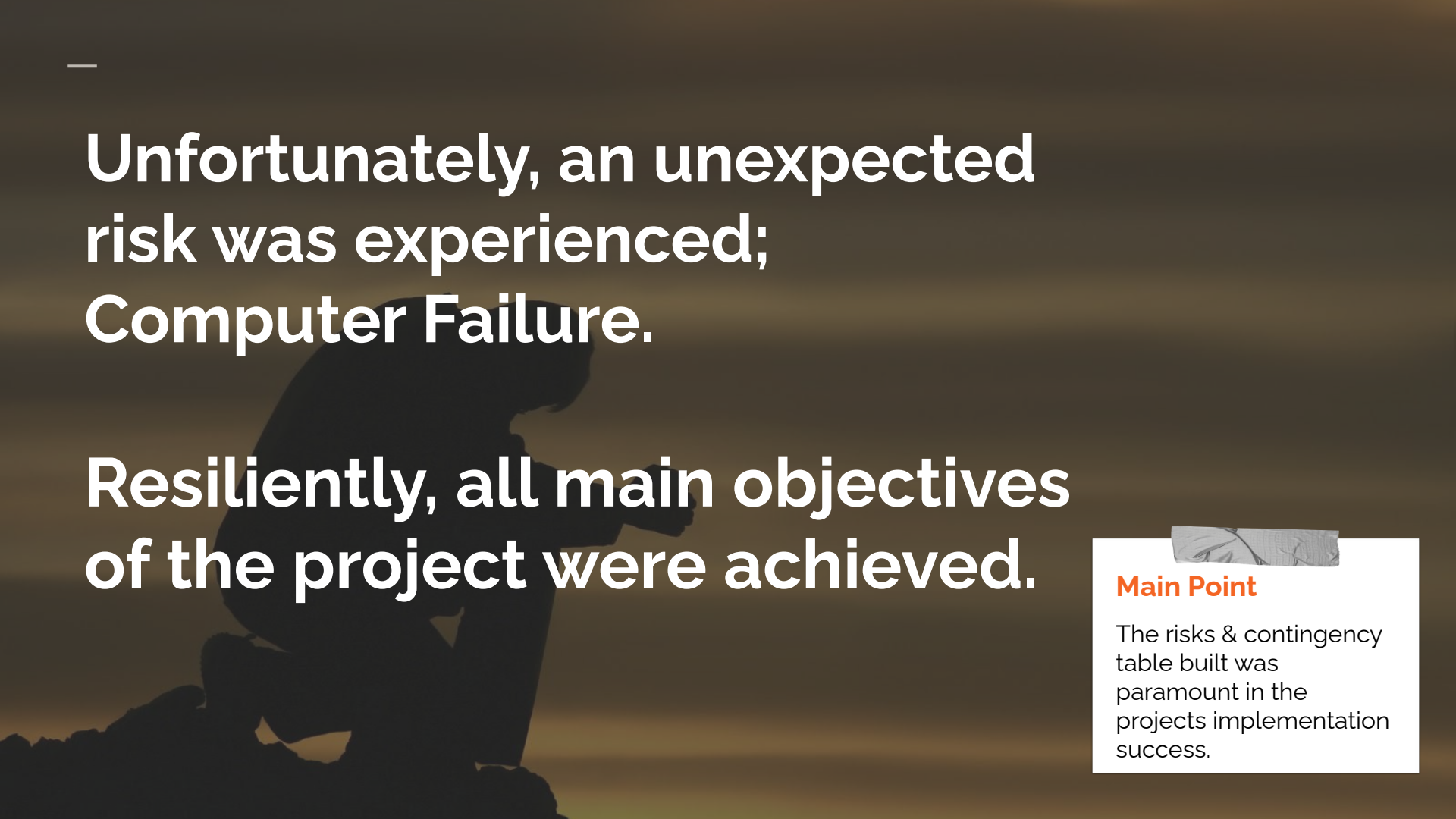
Analyses and visualizes insights using WordClouds.

➔ **Toxic Data Moderation**

Builds text preprocessing, feature extraction and machine learning models to predict hateful sentiment.

Unfortunately, an unexpected risk was experienced; Computer Failure.

Resiliently, all main objectives of the project were achieved.



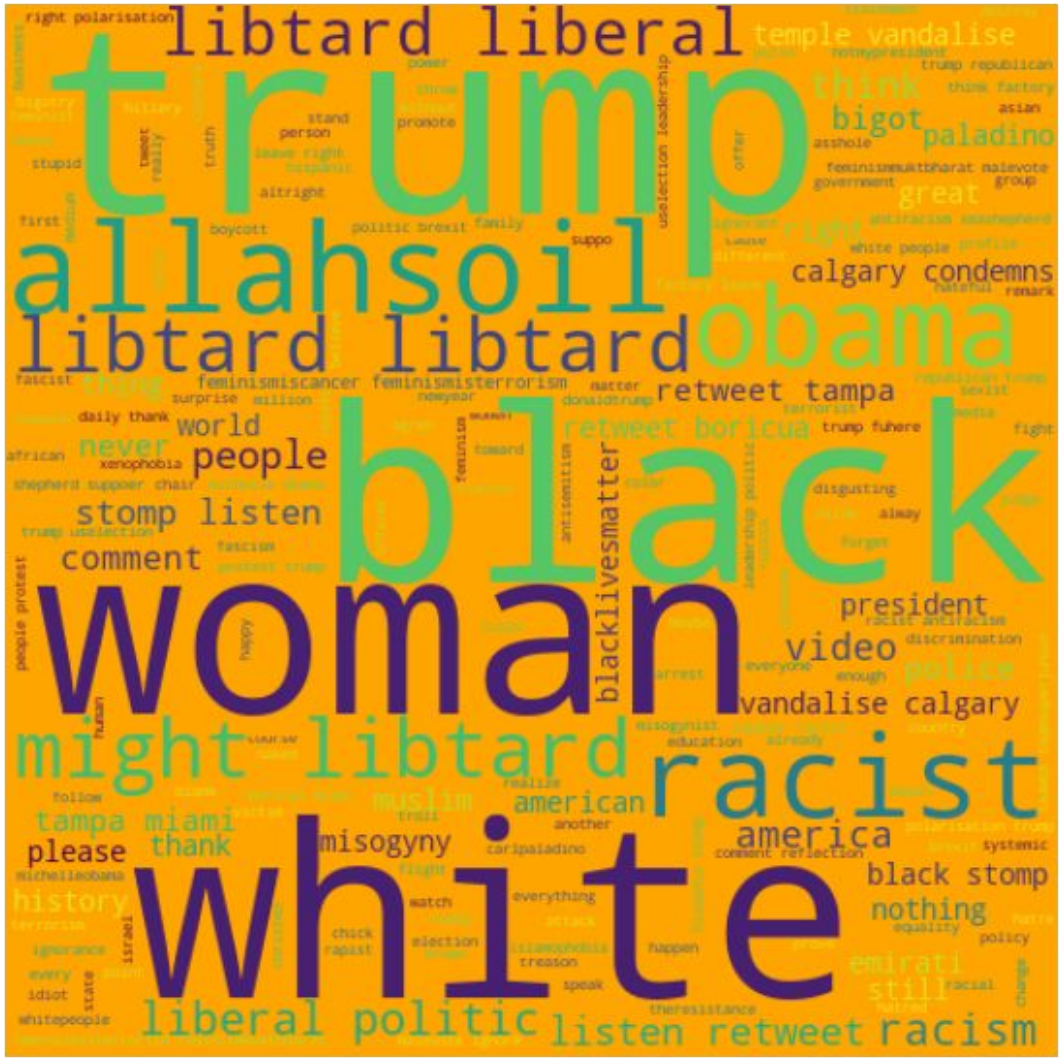
Main Point

The risks & contingency table built was paramount in the projects implementation success.

Objectives Checklist

Specific aims & objectives accomplished during this project:

- ✓ **Gathered data regarding hate speech on the internet.**
- ✓ **Built text preprocessing models for preprocessing toxic data.**
- ✓ **Carried out data analysis on toxic data.**
- ✓ **Generated data visualisations for toxic data .**
- ✓ **Built Bag-Of-Words model for feature extraction.**
- ✓ **Built Word2Vec word embedding model for feature extraction.**
- ✓ **Built Logistic Regression, Support Vector Machine and Neural Network Models for Sentiment Analysis.**
- ✓ **Generated ROC graphs and F-scores for model evaluation.**



Data Analysis.

Data analysis was carried out on our dataset to gain profound insights.

The visual representation on the left were generated using WordClouds. It indicates common toxic words used in conjunction with other words in our dataset.

Data Analysis.

The visual representation on the right were generated using WordClouds. It indicates common non-toxic words used in conjunction with other words in our dataset.

Implementation Results

The model implementation falls into 3 categories with each machine learning algorithm utilising the Bag-Of-Words and Word2Vec models built:

- Logistic Regression Model
- Support Vector Machine Model
- Artificial Neural Network Model

They are each evaluated using AUC and F-score values.

Model 01

Logistic Regression Model

The logistic regression model built using the Bag-Of-Words feature vectors is the baseline for this project implementation. In comparison (on the right) are the AUC and F-score values of the model using Word2Vec feature vectors. Although both provide very similar results, the logistic regression model ran on Word2Vec features was able to categorise more samples correctly.

Area Under Curve

F-Score

Bag-Of-Words

Area Under Curve

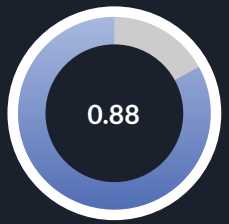
F-Score

Word-2-Vec

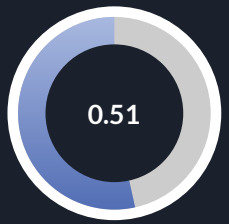
Model 02

Support Vector Machine Model

The Support Vector Model performed marginally similarly to the logistic regression model. Bag-Of-Words vectors are used in comparison (on the right) to the SVM model using Word2Vec feature vectors. SVM performs slightly better when compared to BOW model using the AUC value but worse using F-score values.

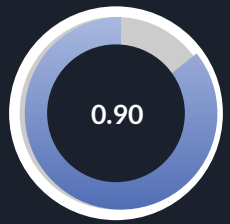


Area Under Curve

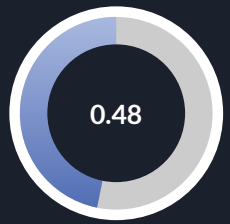


F-Score

Bag-Of-Words



Area Under Curve



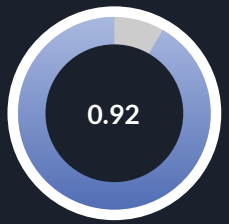
F-Score

Word-2-Vec

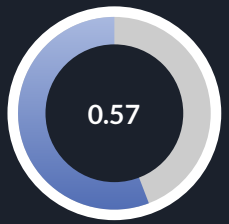
Model 03

Artificial Neural Network Model

The use of an Artificial neural Network is to showcase how advancements have and will improve regulating toxic content on the internet. The highest F-score values are observed using the ANN. The ANN performs better than all previous models. The ANN using Word2Vec vectors performs significantly better compared to using Bag-Of-Words vectors.

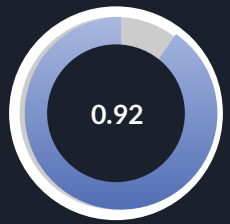


Area Under Curve

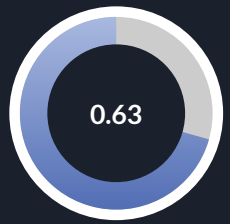


F-Score

Bag-Of-Words



Area Under Curve



F-Score

Word-2-Vec


SOFTWARE
DEMONSTRATION

Project Progress: Current Evaluation

Further Developments


QUESTIONS

Thank you.

REFERENCES

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