**Investigation in using machine learning algorithms to detect if a tweet has a toxic language.**

*Author Saad Iftikhar Project Supervisor Dr Mark Elshaw*

Can machine learning algorithms be used on Twitter to detect if a tweet is toxic?



School of Computing, Engineering and Mathematics,

Coventry University

Computer Science BSc

1st May 2021

# Statement of Originality 6001CEM Declaration of originality

***I Declare that This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.***

# Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialize products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information, please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

# Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Saad Iftikhar Date: 31/12/2021

|  |
| --- |
| First Name: Saad |
| Last Name: Iftikhar |
| Student ID number: 9789180 |
| Ethics Application Number: P130863 |
| 1st Supervisor Name: Mark Elshaw |
| 2nd Supervisor Name |

Table of Content

[Statement of Originality 6001CEM Declaration of originality](#_Toc91856147)

[Statement of copyright](#_Toc91856148)

[Statement of ethical engagement](#_Toc91856149)

[INDEX OF TABLES](#_Toc91856150)

[Index of figures](#_Toc91856151)

[Abstract 1](#_Toc91856152)

[Acknowledgment 2](#_Toc91856153)

[Chapter 1: Introduction 3](#_Toc91856154)

[1.1. Background and Motivation 3](#_Toc91856155)

[1.2. Objectives 3](#_Toc91856156)

[1.3. Structure 3](#_Toc91856157)

[Chapter 2: Literature review 4](#_Toc91856158)

[Chapter 3: Research Methodology 5](#_Toc91856159)

[Chapter 4: Experiments, Evaluation and Results 6](#_Toc91856160)

[Chapter 5: Discussion 7](#_Toc91856161)

[Chapter 6: Project Management 8](#_Toc91856162)

[Chapter 7: Conclusion 9](#_Toc91856163)

[References 10](#_Toc91856164)

[Appendix 11](#_Toc91856165)

# INDEX OF TABLES

**No table of figures entries found.**

# Index of figures

**No table of figures entries found.**

# Abstract

# Acknowledgment

I'd like to express my gratitude to my supervisor, Dr Mark Elshaw, a lecturer in Computer Science at Coventry University, for his unwavering support throughout the research. I was able to enhance my work and complete the dissertation assignment to the best of my ability by offering comments and opinions. I also wanted to express my gratitude to Mark for all of his assistance and encouragement throughout my academic career.

# Chapter 1: Introduction

## Background and Motivation

As mentioned in the project proposal (Iftikhar, 2021). Abusive language is becoming increasingly common in online discussions. It poses a danger to freedom of expression, degrades the dignity of those targeted, and inhibit healthy and beneficial dialogue. Hate speech is not a clear-cut category; it appears to be part of a continuum of discriminatory discourse and is often shown using indirect linguistic means. Words can have similar vocabulary, but the toxicity of the comments is vastly different. Content moderators can't keep up with the flood of user-generated content quickly enough to keep everyone safe. hate-filled content can have considerable psychological risks to a content moderator.

To offer a healthy discussion environment on the Web, automated identification of conflictual languages is required. Hate speech may now be detected more accurately in textual streams because of recent improvements in Natural Language Processing and Natural Language Comprehension. Machine learning algorithms are thought to be the most effective for detecting conflicting languages (Iftikhar, 2021).

In this project, pre-existing tools and algorithms are trained on a dataset and it is evaluated that tool performs best in the detection of abusive language.

Abusive language on an online forum is counter-intuitive and can cause a lot of issues for people of all backgrounds by suppressing their freedom of speech stopping them from expressing themself fully and openly in a positive manner. According to a 2014 Pew Report 4, 73 percent of adult Internet users have witnessed or experienced online harassment, with 40 percent having directly experienced it. Victims of online abuse are frequently from society's most vulnerable groups. Psychological distress, radicalization, and even self-harm and death can result from internet exposure to poisonous and hostile statements (Iftikhar, 2021).

The aim of this project was to make my contribution in the detection of abusive language on any online platform to help reduce this form of oppressive behaviour and help improve these people’s lives (Iftikhar, 2021).

## Objectives

### Model Evaluation

The performance of a model can be measured using Precision, recall, f-measure (f1 score), accuracy and a confusion matrix. The area under the ROC curve (AUC ROC) is a key evaluation metric for classifying toxic comments along with the F1 score and accuracy (Iftikhar, 2021).

### Precision

Precision is also known as the projected value that is positive. It's the percentage of predicted positives that turn out to be true positives (Iftikhar, 2021).

### Recall

The fraction of actual positives that are anticipated positive is known as recall (Iftikhar, 2021).

### F-measure

F-measure is the mean of accuracy and recall. Precision and recall are given equal weight in the standard F-measure (F1) (Iftikhar, 2021).

### Accuracy

Accuracy is the total number of cases that have been successfully categorised (true positives and true negatives) (Iftikhar, 2021).

### Confusion matrix

A confusion matrix shows correctly classified and wrongly classified instances of each class.

The main objectives of this project are as mentioned in the original proposal (Iftikhar, 2021) which are as follows:

* To create the model.
* To review the model.
* It is of high importance that the model has good results on all metrics while keeping the actual positive and actually negative classifications balanced whether they are correctly or incorrectly represented to keep the bias as low as possible.
* To create a model that can detect abusive langue on unseen data with high performance.
* The previous works on the same topic can be used and improved upon in this research.
* This research can be used to provide valuable information for future research.

## Structure

# Chapter 2: Literature review

# Chapter 3: Research Methodology

# Chapter 4: Experiments, Evaluation and Results

# Chapter 5: Discussion

# Chapter 6: Project Management

# Chapter 7: Conclusion

# References

Iftikhar, S. (2021). *Investigation in using machine learning algorithms to detect if a tweet has a toxic language.* Coventry: Coventry Univeristy.

# Appendix