**TEAM MEMBERS DETAILS**

GROUP NAME: Hate Speech Research

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Email | Country | College/Company | Specialisation |
| Kelvin Mpofu | mpofukelvintafadzwa@gmail.com | South Africa | n/a | NLP |

**Problem Description**

The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor. In this problem, We will take you through a hate speech detection model with Machine Learning and Python.

Hate Speech Detection is generally a task of sentiment classification. So for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. So for the task of hate speech detection model, We will use the Twitter tweets to identify tweets containing  Hate speech.

**Business understanding**

Hate speech can have a lot of negative impact and the ability to detect it is very crucial in the age of social media.

**Projected Business advantages**:

* Can improve peoples experience on social media
* Can help alleviate the negative impact on certain groups of people
* Target specific populations more effectively
* Help identify and monitor perpetrators of hate speech

**Project lifecycle**

Diagram

Description automatically generated

The schedule below clarifies the project lifecycle, tasks, and deadlines. The deadlines are split into two, the first draft will be every Friday, and the official one would be on Sundays. Using Github for work and tasks sharings.

|  |  |  |
| --- | --- | --- |
| **Phases** | **Tasks** | **Deadlines** |
| **Ph1.** Understanding the problem | * Data understanding * Business understanding | Week7  1st draft (09/03/2021) Final  (09/05/2021) |
| **Ph2.** Data preparation | * Data cleaning * Type of data * Data problems * Approaches used to overcome the problems in the data * Techniques used | Weeks 8  1st draft (09/10/2021) Final  (09/12/2021)  Week9  1st draft (09/17/2021) Final  (09/19/2021) |
| **Ph3.** Model planning | * Exploratory Data Analysis (EDA) * EDA performed on data * EDA recommendation | Week10  1st draft (09/24/2021) Final  (09/26/2021) |

|  |  |  |
| --- | --- | --- |
|  | - EDA presentation for business users |  |
| **Ph4.** Model building | * Linear model * Ensemble * Boosting | Week11  1st draft (09/01/2021) Final  (09/03/2021) |
| **Ph5.** Communicate results | * Findings are shared with the stakeholders * Ppt presentation | Week12  1st draft (10/08/2021) Final  (10/10/2021) |

**DATA INTAKE REPORT**

Name: NLP – Hate Speech project

Report date: 09/02/2022

Internship Batch: [LISUM02](https://canvas.instructure.com/courses/3110044)

Version:1.0

Data intake by: Hate Speech research

Data intake reviewer:<All group members>

Data storage location: github

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 49 159 |
| **Total number of files** | 2 |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 1,55MB, 3MB |