

**Project:** Hate Speech Detection using Transformers (Deep Learning)

**Week 7:** Deliverables

**Name:** OBIDA ALHAMOUD

**University:** MANISA CELAL BAYAR UNIVERSITY

**Email:** obaida.ismail.alhamoud@gmail.com

**Country:** Turkey

**Specialization:** Data Science

**Batch Code:** LISUM35

**Date:** 17 Aug 2024

**Submitted to:** Data Glacier

**1.Project Lifecycle & Deadlines**

|  |  |  |
| --- | --- | --- |
| **Weeks** | **Date** | **plan** |
| Weeks 07 | May 18, 2022 | Problem Understanding Research hate speech detection techniques. Analyze the problem and define the scope. |
| Weeks 08 | May 25, 2022 | Data Cleaning and Normalization.  Preprocess the tweets by removing noise (e.g., URLs, special characters).  Handle missing data, if any.  Normalize the text data. |
| Weeks 09 | June 1, 2022 | Representation Learning |
| Weeks 10 | June 8, 2022 | Model Building & Training |
| Weeks 11 | June 14, 2022 | Performance Evaluation & Reporting |
| Weeks 12 | June 21, 2022 | Model Deployment & Inference |
| Weeks 13 | June 30, 2022 | Documentation & Submission |

1. **Problem Description** • **Objective:**

o Develop a model to detect hate speech in tweets using deep learning techniques, specifically Transformers.

* + **Definition of Hate Speech:**
    - Any communication that attacks or uses derogatory or discriminatory language against a person or group based on religion, ethnicity, nationality, race, color, ancestry, sex, or other identity factors.
  + **Data Source:**
    - A labeled dataset of tweets where label is 0 or 1 (0 for non-hate speech, 1 for hate speech), and text\_format contains the original tweets.

1. **Business Understanding**

2

* + **Importance of Hate Speech Detection:**
    - Ensures safer online communities by identifying and mitigating hate speech.
    - Supports social media platforms in enforcing policies against harmful content.
  + **Potential Applications:**
    - Content moderation on social media platforms. o Automated reporting of harmful content. o Enhancing user experience by filtering out hate speech.

1. **What type of data do we have:**

A dataset contains 3 features:

1. Id
2. Label
3. Text

Id feature datatype is integer and it contains the tweet Id.

Label is an integer of 0 and 1 and it represents if the text is negative or positive.

Text is a string feature ant it contains the text of tweet.

1. **Approaches to clean the data:**

Using libraries like pandas and re could help us to clean and normalize the dataset

1. **Problems:**

we need to remove special characters and remove all the unnecessary things like:

1. Punctuation
2. URLs
3. @tags

**Punctuation**: it is important to remove the punctuation because is not important.

We remove that using regular expressions.

**URLs:** because we are working on hate speech detection app, we need to give only the text.

**@tags:** we remove @tags using regular expressions

3