**WEEK 8 REPORT**

**Group Name:**

|  |  |  |
| --- | --- | --- |
|  | Member 1 | Member 2 |
| Name | Keilor Fallas Prado | Ky Dang |
| Email | kfallasprado@gmail.com | Keith.dang1610@gmail.com |
| Country | Costa Rica | Vietnam |
| College/ Company |  |  |
| Specialization | NLP | NLP |

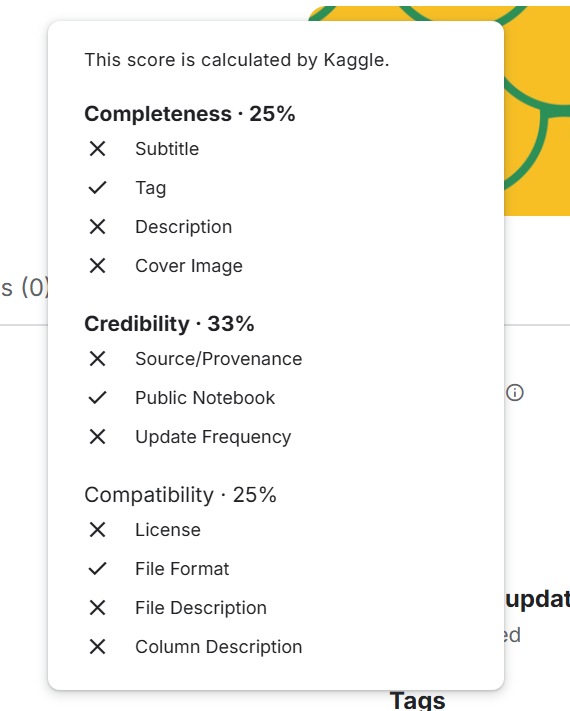
**Problem description:**

Hate speech is a form of communication, whether verbal, written, or behavioral, that attacks or discriminates against an individual or group based on their inherent characteristics, such as religion, ethnicity, nationality, race, gender, or other identity factors. The emergence of hate speech on social media platforms like Twitter poses significant challenges, including creating a toxic environment for users and impacting the platform's reputation.

To address this issue, the goal is to develop a machine learning-based hate speech detection model. This model will classify tweets as hate speech or not, using sentiment classification techniques. By leveraging a dataset of tweets commonly used for sentiment analysis, the model will learn to identify patterns associated with hate speech. This project combines data preprocessing, feature extraction, and machine learning to build an effective classifier capable of addressing the problem of online hate speech in a scalable and automated way.

**Data understanding:**

Due to the dataset was collected from the twitter, it’s outdated, and the usability score is quite low 2.94.  
- The data was collected six years ago and has not been regularly updated, rendering it obsolete.



The dataset has two subdatasets: train (31962 records) and test (17197 records). The train dataset includes: 3 columns (id, label, tweets).

I identified inconsistencies in the train dataset where the labels 0 and 1 do not align with the content of the tweets. For instance, tweets that are likely hate speech are labeled as 0, and vice versa

**Project lifecycle**

|  |  |  |
| --- | --- | --- |
| **Weeks** | **Due date** | **Plan** |
| Week 8 | 11/26/2024 | Review data source and ensure it is representative of hate speech contexts. |
| Week 9 | 12/02/2024 | Remove duplicates, nulls, and irrelevant data. |
| Week 10 | 12/09/2024 | Evaluate and select models such as Logistic Regression, SVM, or Transformers (e.g., BERT). |
| Week 11 | 12/16/2024 | Tokenization - Identify relevant linguistic and contextual features |
| Week 12 | 12/23/2024 | Training and evaluation model |
| Week 13 | 12/30/2024 | Document the challenge |

**Github Repo link:**

* Individual GitHub links:
  + KyDang: <https://github.com/KeithDang1610/NLP_HateSpeech-Detection>
  + Keilor: