# WEEK 12 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:**

In the previous week, we experimented with two tweet featurization methods. The first method involved manually extracting features from the tweet data, which allows us to use simpler machine learning or simple deep learning models for classification. This approach is more efficient in terms of processing time but may not capture the complex patterns and contextual relationships in the data as effectively as deep learning models. The second method uses **DistilBERT** for tokenization, which automatically processes the text with a pre-trained deep learning model. While this method takes longer due to the complexity of the model, it benefits from capturing more nuanced information and contextual relationships within the text. The goal is to determine whether the additional computational cost of DistilBERT provides a significant improvement in hate speech detection compared to the simpler, faster manual feature extraction method.

We decided to train a model using the second method, then evaluate its performance and finally predict the labels for the test dataset.

**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: