Project Report

Team Member's Details

Group Name: Scalable Minds
Members:

- **Name:** Lucy Nowacki and potential other participants
  Email: quantlucy@gmail.com
  Country: UK
  College/Company: Queen Mary
  Specialization: NLP

Problem Description

Technolgy

PyTorch GPU, Azure

Introduction

In this project, we aim to tackle the problem of hate speech detection using various machine learning approaches. Specifically, we will implement and compare two models:

1. **NanoGPT**
2. **xLSTM (Extended Long Short-Term Memory)**

We will train these models and compare their performance against state-of-the-art pre-trained models such as GPT, BERT, and ILAMA.

Objective

The primary objective is to develop lightweight models that require significantly less memory and computational power, making them suitable for deployment on devices with limited resources like smartphones and tablets.

Business Understanding

Background

Hate speech on social media platforms and other online forums is a growing concern. Detecting and mitigating such harmful content is crucial to maintaining a safe and inclusive digital environment. Traditional large language models (LLMs) like GPT, BERT, and ILAMA are highly effective but require

significant computational resources, making them impractical for deployment on resource-constrained devices.

Traditional large language models (LLMs) like GPT, BERT, and ILAMA are highly effective but require significant computational resources, making them impractical for deployment on resource-constrained devices.

Goal

Our goal is to create efficient and lightweight models for hate speech detection that can be deployed on low-resource devices. This will enable real-time detection and filtering of harmful content without the need for powerful computational infrastructure.

In addition to efficiency, another aim is to build models with bespoke architectures that are relatively simple. These models should be easy to adjust to different platforms and scalable through parallelization. This flexibility will ensure that the models can be adapted for various deployment environments, enhancing their utility and longevity.

For future development beyond this project, the model should also be designed to accommodate not only language data but also visual data enriched by physical reality. This extension will involve integrating neural operators to handle continuous data and complex dependencies across different data modalities. By doing so, we aim to create a comprehensive solution capable of processing and understanding multi-modal data inputs, paving the way for more sophisticated applications.

Benefits

- **Accessibility:** Enable hate speech detection on a wider range of devices, including smartphones and tablets.
- **Cost-Efficiency:** Reduce the need for expensive computational resources, making the technology more accessible to smaller organizations and developers.
- **Scalability:** Facilitate the deployment of models in resource-constrained environments, allowing for wider adoption and impact.
- **Real-Time Processing:** Allow for the real-time detection and mitigation of hate speech, enhancing the user experience and safety on digital platforms.
- **Flexibility:** The bespoke architecture ensures the model can be easily adjusted to different platforms, improving adaptability and deployment efficiency.
- **Future-Proofing:** The capability to integrate neural operators for handling both language and visual data ensures the model remains relevant and expandable for future applications involving multi-modal data inputs.

Project Lifecycle Along with Deadline

Phases and Timeline

3. **Problem Definition and Business Understanding (May 16 - May 20)**
    - Define the problem.
    - Understand the business context and objectives.
    - Generate a data intake report.
4. **Data Collection and Preparation (May 21 - May 31)**
    - Collect and preprocess hate speech data.
    - EDA and data interpretation
    - Detection of problems in the data (number of NA values, outliers , skewed etc) and how to overcome them.
5. **Model Development (June 1 - June 15)**
    - Implement NanoGPT and xLSTM models.
    - Train each model using the prepared dataset.
6. **Evaluation and Comparison (June 16 - June 25)**
    - Evaluate the models using key performance metrics.
    - Compare the results with state-of-the-art pre-trained models (GPT, BERT, ILAMA).
7. **Documentation and Final Submission (June 26 - June 30)**
    - Compile the final report.
    - Submit the project by July 1.

Deadline

The final project submission deadline is **July 1st**.

Important !!!

Separate reports and deliverables will be provided according to the specified schedule of LISUM32.

DATA INTAKE REPORT

Introduction

The Twitter Hate Speech dataset, hosted on Kaggle, includes tweets labeled for hate speech and offensive language. This dataset is essential for developing machine learning models aimed at detecting and mitigating hate speech on social media platforms. It is divided into training and test datasets, each containing different columns of information.

Data Sources

The dataset comprises two CSV files:

8. **Training Data**: **train_E6oV3lV.csv**
9. **Test Data**: **test_tweets_anuFYb8.csv**

- **Dataset URL**: Twitter Hate Speech on Kaggle

Dataset Overview

Training Data (train_E6oV3lV.csv):

- **Volume**: 31,962 records
- **Columns**: 2 columns
    - **id**: Unique identifier for each tweet
    - **label**: Classification label
        - 0: Hate Speech
        - 1: Offensive Language
    - **tweet**: Text of the tweet

Test Data (test_tweets_anuFYb8.csv):

- **Volume**: 17,197 records
- **Columns**: 1 column
    - **id**: Unique identifier for each tweet
    - **tweet**: Text of the tweet

Data Structure and Characteristics

Training Data Details:

- **id**: Numerical identifier ranging from 1 to 31,962.
    - No missing or mismatched values.
- **label**: Categorical values with 0 (Hate Speech) and 1 (Offensive Language).
    - No missing or mismatched values.
    - Distribution:

train_df['label'].value_counts()

        - 0: 29,720 tweets
        - 1: 2,242 tweets

- **tweet**: Text data containing the content of the tweet.
    - 29,530 unique values, indicating some tweets are repeated.

Test Data Details:

- **id**: Numerical identifier ranging from 31,963 to 49,159.
    - No missing or mismatched values.
- **tweet**: Text data containing the content of the tweet.
    - 16,130 unique values, indicating some tweets are repeated.

Data Quality and Integrity

- **Completeness**: Both datasets are complete with no missing values in any columns.
- **Uniqueness**: The **tweet** column has several unique values, but some repetition is present. Each **id** is unique in its respective dataset.
- **Accuracy**: Labels are manually annotated, which is generally accurate, though subjective bias may be present.
- **Consistency**: The datasets are consistent in formatting and data types.

Potential Data Usage

The datasets are primarily used for:

- **Text Classification**: Building models to classify tweets into hate speech or offensive language.
- **Natural Language Processing (NLP)**: Preprocessing tasks such as tokenization, stemming, and lemmatization.
- **Sentiment Analysis**: Understanding the sentiment expressed in tweets.
- **Feature Analysis**: Examining the distribution and significance of different features across the labels.

Conclusion

The Twitter Hate Speech dataset is well-structured and suitable for text classification tasks. By leveraging this dataset, models can be trained to detect harmful content on social media platforms effectively. The preprocessing steps outlined ensure that the data is clean and ready for analysis, enabling robust model development and evaluation.