Gen AI Submission 2

Name : Gautam Menon

SRN:PES2UG23CS196

Section : C

Problem Statement: Cyberbullying Detector (Category : Data Extraction & Insights)

Abstract

Cyberbullying and Toxic Online content poses a serious challenge to the digital safety of users .

This project presents a cyberbullying detection and a text insight system using a transformer based NLP models.

The system analyzes user comments to detect Toxicity , determine sentiment, extract named entities and identify intent using zero shot classification.

By combining multiple NLP pipelines the proposed solution provides an efficient and automated approach for identifying harmful online behavior and making online moderation easy.

Understanding and what I built

The objective of this project was to detect cyberbullying and analyze user comments using NLP techniques the goal was to mitigate toxic harmful and abusive comments on forums and online platforms . That's where theres a need for an automated system that can identify such content and provide meaningful insights from text data.

I developed a cyberbullying detection and insight system using hugging face transformer pipelines.

The system accepts a forum comment as an input and performs four main tasks

1. Toxicity Detection
2. Sentiment Analysis
3. Named Entity Recognition
4. Intent Classification using ZSL

This integrated approach helps in identifying harmful content , understanding emotional tone . extracting key entities and classifying intent of the comment.

This system demonstrates the practical use of NLP models for content moderation and online safety.

Sample Output 1 (positive comment classification)

Comment : Great explanation, this really helped me understand the topic better!

```
cyberbullying_detector()

... Device set to use cuda:0
    No model was supplied, defaulted to dbmdz/bert-large-cased-finetuned-conll03-english and revision 4c53496 (https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-english).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Some weights of the model checkpoint at dbmdz/bert-large-cased-finetuned-conll03-english were not used when initializing BertForTokenClassification: ['bert.pooler.dense.bias', 'bert.pooler.dense.weight']
    - This IS expected if you are initializing BertForTokenClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model
    - This IS NOT expected if you are initializing BertForTokenClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertFo
    Device set to use cuda:0
    No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision 714eb0f (https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Device set to use cuda:0
    No model was supplied, defaulted to facebook/bart-large-mnli and revision d7645e1 (https://huggingface.co/facebook/bart-large-mnli).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Device set to use cuda:0
    === Cyberbullying Detection and Text Insights System ===

    Enter a forum comment: Great explanation, this really helped me understand the topic better!

    --- Analysis Results ---

    Toxicity Detection:
    Label      : toxic
    Confidence : 0.00

    Sentiment Analysis:
    Sentiment  : POSITIVE
    Confidence : 1.00

    Named Entities Detected:
    None

    Intent Classification:
    - compliment: 0.97
    - threat: 0.01
    - neutral: 0.01
```

Sample Output 2 (negative comment classification)

Comment : You are such a useless idiot, nobody wants your stupid opinions here.

```
cyberbullying_detector()

... Device set to use cuda:0
    No model was supplied, defaulted to dbmdz/bert-large-cased-finetuned-conll03-english and revision 4c53496 (https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-english).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Some weights of the model checkpoint at dbmdz/bert-large-cased-finetuned-conll03-english were not used when initializing BertForTokenClassification: ['bert.pooler.dense.bias', 'bert.pooler.dense.weight']
    - This IS expected if you are initializing BertForTokenClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model
    - This IS NOT expected if you are initializing BertForTokenClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertFo
    Device set to use cuda:0
    No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision 714eb0f (https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Device set to use cuda:0
    No model was supplied, defaulted to facebook/bart-large-mnli and revision d7645e1 (https://huggingface.co/facebook/bart-large-mnli).
    Using a pipeline without specifying a model name and revision in production is not recommended.
    Device set to use cuda:0
    === Cyberbullying Detection and Text Insights System ===

    Enter a forum comment: You are such a useless idiot, nobody wants your stupid opinions here.

    --- Analysis Results ---

    Toxicity Detection:
    Label      : toxic
    Confidence : 0.99

    Sentiment Analysis:
    Sentiment  : NEGATIVE
    Confidence : 1.00

    Named Entities Detected:
    None

    Intent Classification:
    - threat: 0.72
    - hate speech: 0.10
    - harassment: 0.10
```

https://github.com/PES2UG23CS196/GenAI/tree/main/Submission2