

ML - Hackathon Set 3 (Sentiment Analysis)

UE22CS352A

1. Overview

The Quest for the Perfect Fusion Model:

In the bustling city of Techville, where innovation and creativity were celebrated, there was a team of bright minds known as the Gyaan Yoddhas. Led by Professor Mitesh Khapra, a renowned expert in artificial intelligence, the team comprised:

- Manisha Bantia: The video analysis wizard, skilled in extracting meaningful patterns from visual & audio data.
- Avisek Lahiri: The natural language processing (NLP) guru, adept at deciphering the intricacies of textual information.
- Anand Mishra: The data engineer, responsible for collecting and preprocessing vast amounts of data.
- Prachi Jain: The machine learning specialist, who could build and fine-tune complex models with ease.

One sunny afternoon, as the Gyaan Yoddhas gathered in their lab, Professor Mitesh unveiled their next grand challenge.

"Team, we have been tasked with a monumental project" Professor Mitesh announced. "Our mission is to develop a multi-modal fusion model for sentiment analysis using video and text data from a TV show: Friends"

Anand looked intrigued. "Friends? That is an iconic show! What is the specific goal?"

Professor Mitesh smiled. "For the videos given, we need to detect the **sentiments**. The system should classify the clips into categories based on how the speaker feels when they speak. It should also identify the sentiment of each sentence spoken. We'll be using both **video** and **subtitle** data to analyze the evolution of sentiment over time." We will work on this challenge to develop a robust multi-modal model.

2. Description

You are provided with a dataset consisting of video and text data, along with their corresponding sentiment labels.

Make a team with 4 members in this competition. Team name should follow the below naming convention:

PESU_{campus}_{srn1}_{srn2}_{srn3}_{srn4} using the last 3 digits of the student's SRN.

Eg: PESU_EC_546_542_444_355

Note - Please refer to the same team name for all submissions of deliverables.

3. Deliverables

There are 3 deliverables:

- a. Make a submission in the competition. Refer Submission Format. You can make up to 5 submissions and you can choose which submission should be counted for the final evaluation (only 1, by default Kaggle takes the submission with highest score). The primary metric for evaluation is **Accuracy**.
- b. Submit the final .ipynb file in the google form shared. Ensure that your notebook is comprehensive and includes relevant code and output.
- c. A PDF document containing the following:
 - i. **Data Preprocessing Steps** - Outline the steps taken to clean and prepare the dataset for model training.
 - ii. **Feature Extraction Steps** - Detail the features selected or engineered to enhance model performance.
 - iii. **Early vs Late Fusion** - Provide a comparative analysis of any fusion techniques used, with insights into their effectiveness.
 - iv. **Model Decisions** - Explain the reasoning behind choosing specific models or methodologies.
 - v. **Performance Analysis** - Discuss model performance, including key metrics and any insights gained from experimentation.

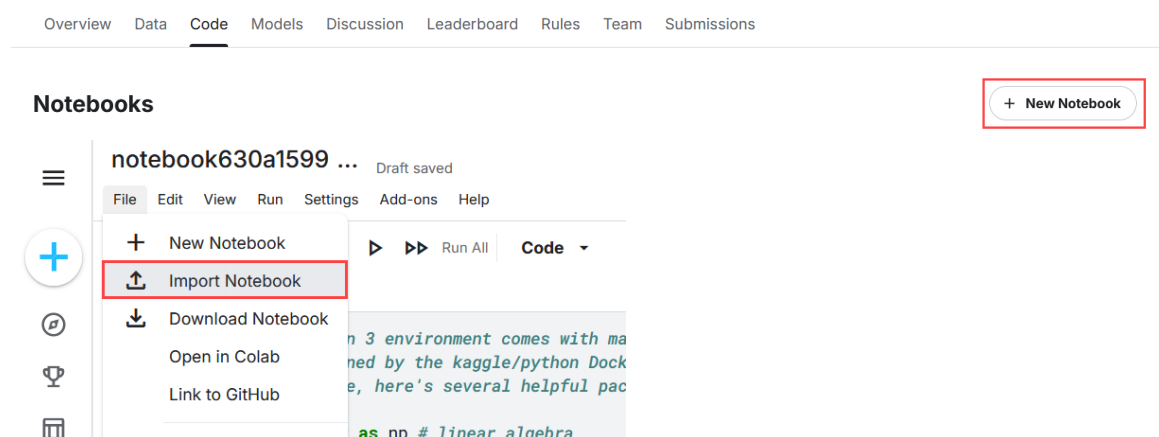
Make sure the .ipynb file and the .pdf file are named the same as your team name (Eg: PESU_EC_546_542_444_355.ipynb).

The .pdf file should also have the Name and SRN of the 4 team members in the beginning of the document.

4. Submission Format

Submissions are to be made **ONLY** through a Kaggle notebook. Direct submission of a .csv file will not be considered towards the final evaluation.

In case you are using your own coding environments, create a notebook under the code section of the competition and import your file.



The notebook submitted should finally generate a `submission.csv` file. The file should have two columns: ``Sr No.`` and ``Sentiment``. Failure to follow the above syntax **will cause an ERROR in submission** and a loss of one of your 5 attempts.

Example code:

```
submission_df = pd.DataFrame({
    'Sr No.': test_ids,
    'Sentiment': model_predictions
})
submission_df.to_csv("submission.csv", index=False)
```

Make sure the file name is exactly `submission.csv` and not `submissions.csv` or anything else.

The values in ``Sr No.`` should correspond to the values of ``Sr No.`` in the `test/text.csv` file. The ``Sentiment`` values should be the predicted values of your model.

To submit, go to the notebook you want to submit, and on the right side under the heading **Submit to competition** click on the **Submit** button.


Notebook

Input

Output

Table of contents

Submit to competition

 ML Hackathon EC Campus Set 2

LATEST SCORE

BEST SCORE

DAILY SUBMISSIONS

-

-

3 / 5 used

Submit

:

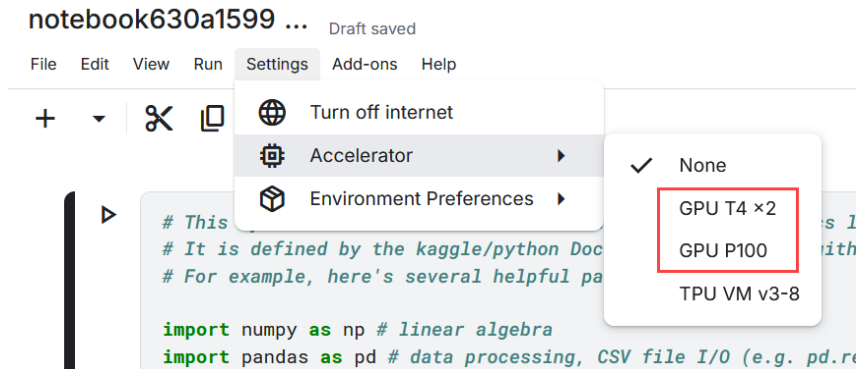
Session options

An example subset of `submission.csv` is shown below:

```
Sr No.,Sentiment
61,positive
72,negative
103,neutral
.
.
.
```

5. Miscellaneous Instructions

- Be careful enough not to make the notebooks public.
- Make sure to use accelerators to run the notebook. (T4 & P100 GPU's)



- Use only the dataset provided in Kaggle. No external datasets are allowed to be used for training / testing.
- Using API Keys & Pre-trained LLMs are strictly forbidden, however embeddings such as GloVe & Word2Vec can be used.
- The following tools can be used:
 - Pytorch
 - Sci-kit Learn
 - Keras
 - Tensorflow

6. Dataset Description

You have a multi-modal dataset created to support sentiment recognition in conversational contexts. Sourced from the TV show Friends, it provides visual and text data, enabling comprehensive analysis of sentiments within context.

There are two folders, /train and /test.

Under each folder, there is a text.csv and a /videos folder.

The text.csv has the following format in the /train folder.

Column Name	Description
Sr No.	Unique number referring the utterance
Utterance	Individual utterances as a string.
Speaker	Name of the speaker associated with the utterance.
Sentiment	The sentiment expressed by the speaker in the utterance.
Dialogue_ID	The ID of the Dialogue
Utterance_ID	The ID of the Utterance in the Dialogue
Season	The season no. of Friends TV Show to which a particular utterance belongs

Column Name	Description
Episode	The episode no. of Friends TV Show in a particular season to which the utterance belongs.
StartTime	The starting time of the utterance in the given episode in the format 'hh:mm:ss,ms'.
EndTime	The ending time of the utterance in the given episode in the format 'hh:mm:ss,ms'.

The `text.csv` under the `/test` folder has all the columns above except ``Sentiment`` which is the target.

There are 3 possible labels for sentiment:

- neutral
- positive
- negative

Each utterance in a dialogue has been labelled by sentiment annotation and the corresponding video file can be inferred using the ``Dialogue_ID`` and ``Utterance_ID``. Videos have a naming convention of `dia{Dialogue_id}_utt{Utterance_id}.mp4`.

There are 1000 training samples and 100 test samples.

There is a possibility while trying to read the `.csv` file, you run into errors of invalid bytes in some locations.

In this case, try using a different encoding like `ISO8859-1` and `CP-1252`. Refer [here](#) for other encodings.

6. Competition Rules

The following rules are established to ensure a fair and ethical Kaggle competition. Participants are expected to comply with these rules throughout the competition.

1. Naming Convention of Team Names:

The students should form a team in each class. (Team size is 4)

`PESU_{campus}_{srn1}_{srn2}_{srn3}_{srn4}` (Use only last 3 digits of SRN).

Please refer to the same team name for all submissions.

2. Data Usage:

Participants must use the provided dataset for training their models. Using external datasets without permission is prohibited and will result in disqualification or penalties.

3. No usage of LLM API's:

Pretrained LLM models are not allowed. Participants should build their models from scratch, including data extraction, pre-processing, neural network architectures, training and evaluation. Usage of API Keys are **strictly prohibited** and will cause an instant disqualification of the team. However, embeddings like GloVe and Word2Vec may be used.

4. **Plagiarism:**
Plagiarism is strictly prohibited. Participants should submit their own work and provide proper citations if they use external resources or references. If plagiarism is found between 2 notebooks, your team will be disqualified.
5. **Code Sharing:**
Sharing code and models among participants during the competition is discouraged. Each participant is expected to develop their solution independently.
6. **Submission Format:**
Submissions must follow the specified format and naming conventions for files. Failure to do so may result in submission issues. You must submit as a notebook and not just upload a CSV file.
7. **Documentation:**
Participants should maintain clear and well-documented code and comments in their notebooks. Documentation helps reviewers and fellow participants understand the methodology.
8. **Communication:**
Respectful and constructive communication is expected among participants in Kaggle forums and discussions. Harassment or abusive behaviour is not tolerated.
9. **Review Process:**
The competition will undergo a review process for rule compliance. Organizers reserve the right to disqualify participants who violate the rules or engage in unethical practices.