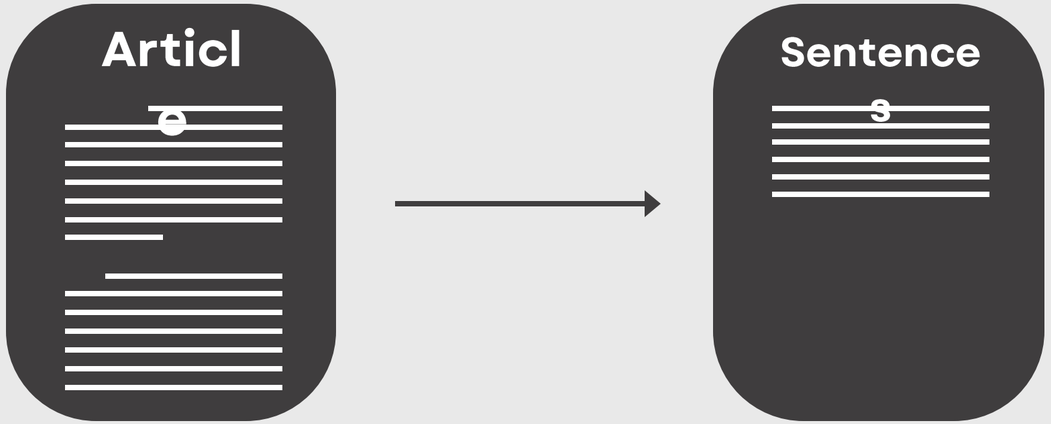
INTERNSHIP REPORT

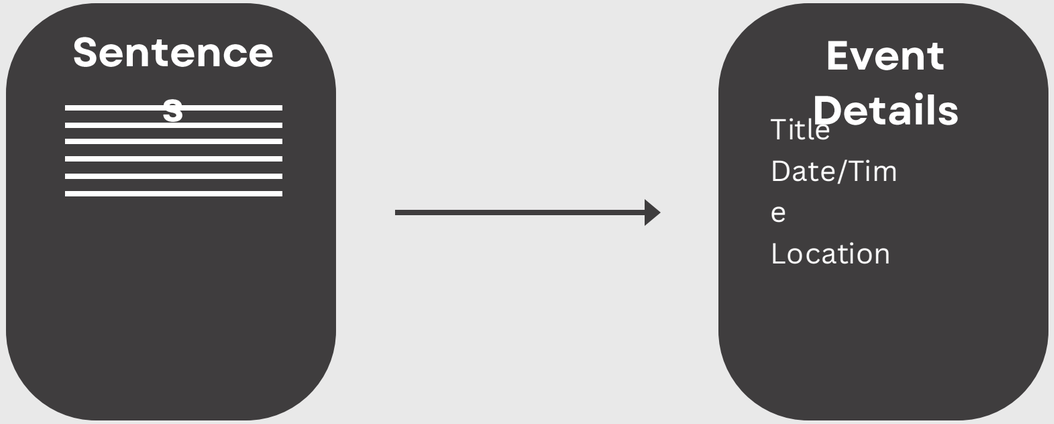
By Sai Pranay Deep

PROJECT GOALS

1. Event Information sentences Extraction from Long-form Textual Data(monologues)



2. Extract the Event Information from the sentences extracted above



PROBLEM STATEMENT - 1

We have long-form articles or monologues that need to be reduced into a few sentences, making them easier to input into a large language model (LLM) for further processing.

Sample Input:

The Tech Summit is going to happen. Come and Join. It is at Cafeteria. We will have fun.

The Tech Summit is going to happen. → Title present

Come and Join. → No useful information

It is at Cafeteria. → Location present

We will have fun. → No useful information

Sample Output:

The Tech Summit is going to happen. It is at Cafeteria.

Approach - 1 : Semantic Search

We will use a set of base sentences representing key information such as the title, date/time, and location. Each sentence in the article will be compared to these base sentences using a cosine similarity. If the average similarity score of a sentence exceeds a predefined threshold, the sentence will be retained; otherwise, it will be discarded. This approach helps filter out irrelevant or less important content, focusing only on sentences closely aligned with the core information.

Sample Base Sentences:

Title

- announces the event
- invited to the party
- hosting the event

Location

- Location of the event is delhi
- Where it happens?
- location
- address
- venue

Date/Time

- Date of event
- Time of program
- will held on 27th August 2025
- on Thursday morning
- at 9 o clock
- starts at 10 AM

Cosine Similarity

Date of the program is 01 Jan 2004
The Tech Summit is going to happen on 17th July

0.313

Date of the program is 01 Jan 2004
Don’t miss this

0.05

Announces the event
You are all welcomed to Tech Expo

0.393

Date of the program is 01 Jan 2004
Please have dinner

0.09

Cosine Similarity Evaluation

Calculate Cosine Similarities

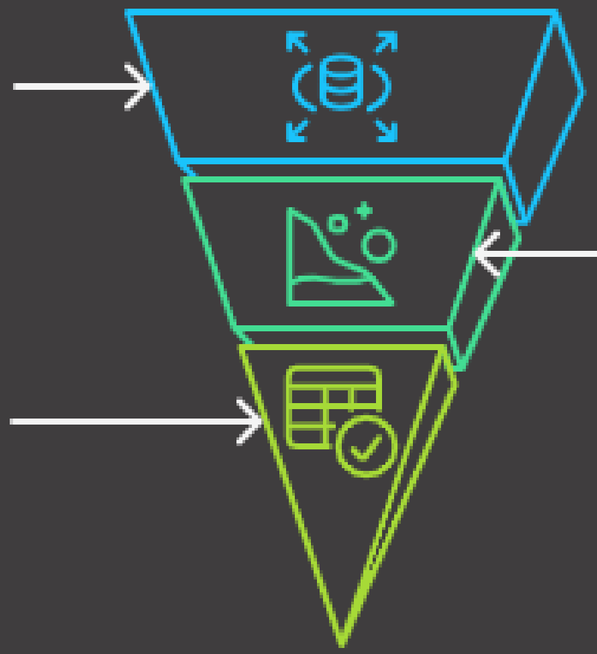
Computing similarity scores between sentences

Average Similarities

Finding the average of similarity scores

Check Threshold

Comparing average score with predefined threshold



Accuracy Calculation

- Let required output should contains sentences (1, 2, 5, 6, 8, 10)

(1, 2, 5, 6, 7, 8, 10, 11) ✓ (1, 5, 6, 7, 8, 10, 11) X 2 is missing

- If the predicted output contains all required sentences, then prediction is correct.
- Otherwise prediction is wrong
- Accuracy = correct predictions / Total samples

Model Comparision

	MiniLM	LaBSE
Accuracy	• ~82% accuracy with noise • ~5-10 extra noise sentences	• ~84% accuracy with noise • ~5-10 extra noise sentences
Size	• ~27M parameters • ~80MB	• ~470M parameters • ~1.8GB
Speed	• Fast	• Slow

Approach - 2 : Fine-Tuning a Binary Classification Model

We build a database of event-related sentences. Each sentence is evaluated to determine whether it contains key event information such as a title, date/time, and location.

- Sentences that include any three components (title, date/time, and location) are labeled with a 1.
- Sentences that are missing all of these components are labeled with a 0.

Using this labeled dataset, we fine-tune a binary classification model to automatically identify sentences that contain complete event information.

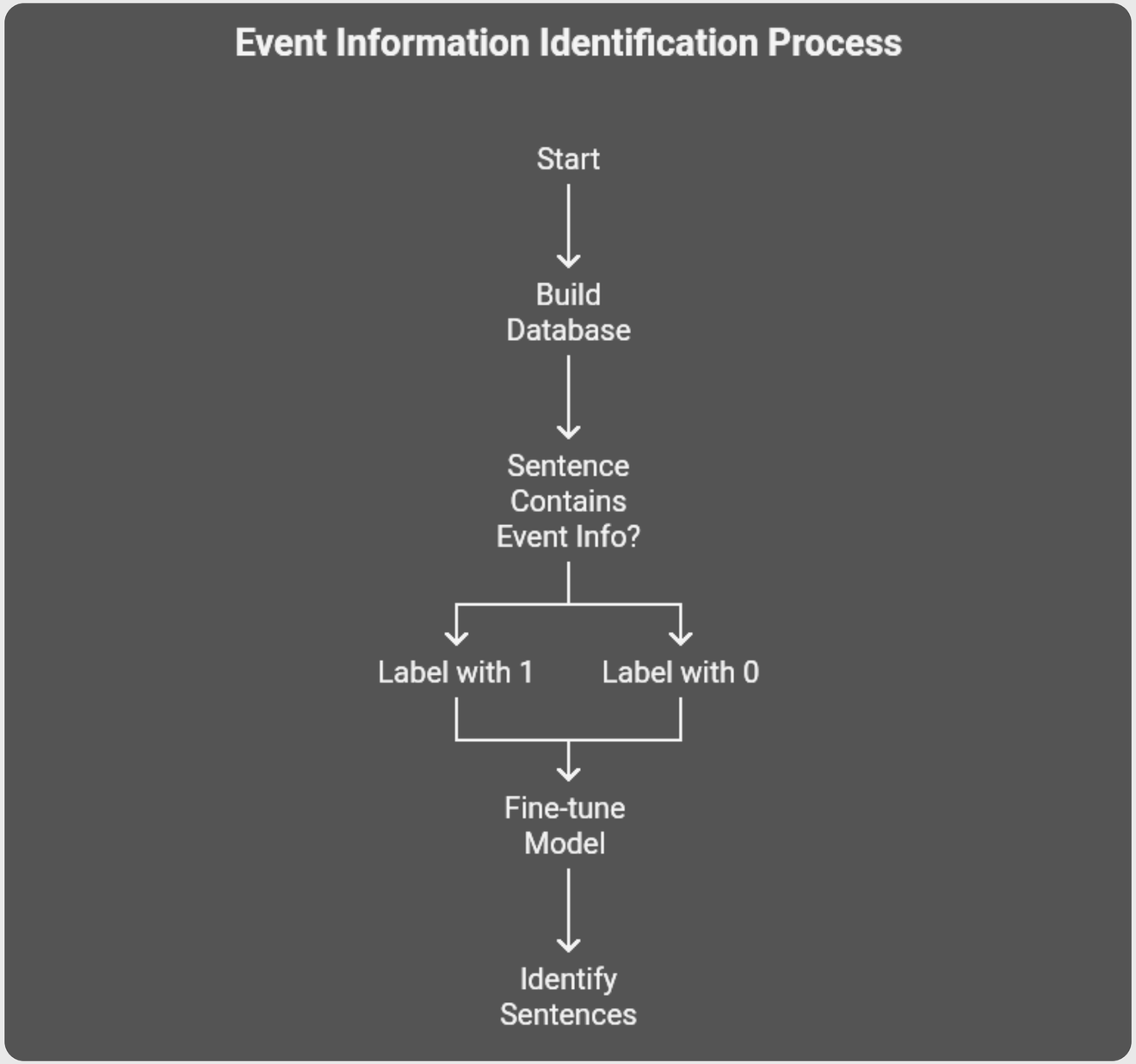
Sample Sentences:

The Startup Pitch starts tomorrow at 11 AM, 1
- 1 as it contains title and time

Catch the Orientation Day live at Innovation Hub, 1
- 1 as it contains title and location

A celebration was traditional, 0
- 0 as there is no title, date/time, location

The ceremony proceeded smoothly with all participants engaged, 0
- 0 as there is no title, date/time, location


Event Information Identification Process
Start
Build Database
Sentence Contains Event Info?
Label with 1
Label with 0
Fine-tune Model
Identify Sentences

Dataset Used for fine-tuning a binary classifier

Prepared by	**Ollama**
No. of samples	**1069**
Event info related	**574**
Non Event related	**495**

Approach - 2 : Fine-Tuning a Multi-label Classification Model

We build a database of event-related sentences. Each sentence is evaluated to determine whether it contains key event information such as a title, date/time, and location.

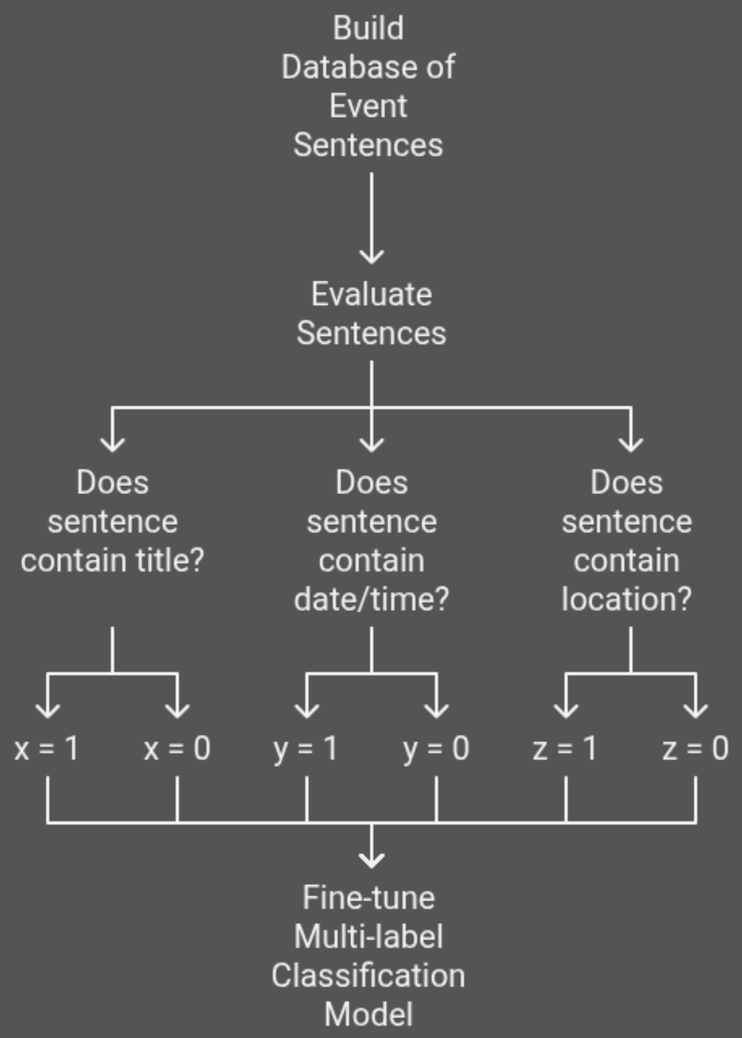
- label = (x, y, z)
- x = 1 if sentence contains title related information. Otherwise x = 0
- y = 1 if sentence contains date/time related information. Otherwise y = 0
- z = 1 if sentence contains location related information. Otherwise z = 0

Using this labeled dataset, we fine-tune a multi-label classification model to automatically identify sentences that contain complete event information.

Sample Sentences:

- "Don't miss the Winter Wonderland on January 3rd, 2026 at Snow Park." - (1,1,1)
- The Hip-Hop Poetry Night happens at The Urban Beat Café. -(1,0,1)
- Be there for the introduction of our Game-Changing Platform. - (1,0,0)
- "Brace for the Book Launch - happening on 29 August 2025, 7:45 AM." - (1,1,0)
- Final segment concludes at 5:45 PM Friday. - (0,1,0)
- Be there at Bengaluru Tech Park on 9:45 AM. - (0,1,1)
- Don't miss us at Club Escape. - (0,0,1)
- A demonstration changed perceptions. - (0,0,0)

Event Information Identification Process



Dataset Used for fine-tuning a multi-label classifier

Prepared by	Ollama
No. of samples	6411
Only Title containing sentences	652
Only Date/Time sentences	643
Only Location sentences	455
Both Title, Date/Time sentences	674
Both Title, Location sentences	960
Both Date/Time, Location sentences	520
All Title, Date/Time, Location sentences	810
No related sentences	1697

Approaches Comparision

Semantic Search	Binary Classifier	Multi-label Classifier
~84% accuracy excluding noise	~90% accuracy excluding noise	~96% accuracy excluding noise
Model size cannot be reduced	Model size can be reduced using quantization & distillation	Model size can be reduced using quantization & distillation
No training time required	1hr+ training time required	1hr+ training time required
No training data required	Training data required	Training data required
Output is automatically seperated into categories	Output should be separated into categories	Output is automatically seperated into categories
5-10+ noise sentences	5-10 noise sentences	3-7 noise sentences

PROBLEM STATEMENT - 2

We need to extract key event-related information – title, date/time, and location – from the sentences identified in the previous step.

Sample Input:

- Be there for the debut of our Industry-Disrupting Technology.
- "Mark your calendar for the Coding Bootcamp on 30 September 2025, 11:45 PM."

Sample Output:

- Title : Industry-Disrupting Technology
- Title : Coding Bootcamp
  Date/Time : 30 September 2025, 11:45PM

Approach - 1 : Using Libraries

To Extract Time:
There are several libraries available for extracting date and time information from text. One of the most effective and widely used options is **parsedatetime**.
However, it does not perform well on certain edge cases. To address this, we incorporated additional fuzzy matching terms and custom handling logic, which significantly improved accuracy in those cases.

To Extract Location:
There are few libraries available for extracting location from text. But they can only extract locations like Countries, Cities but not general locations.

To Extract Title:
There are few libraries available for extracting location from text. But they can only extract locations like Countries, Cities but not general locations.

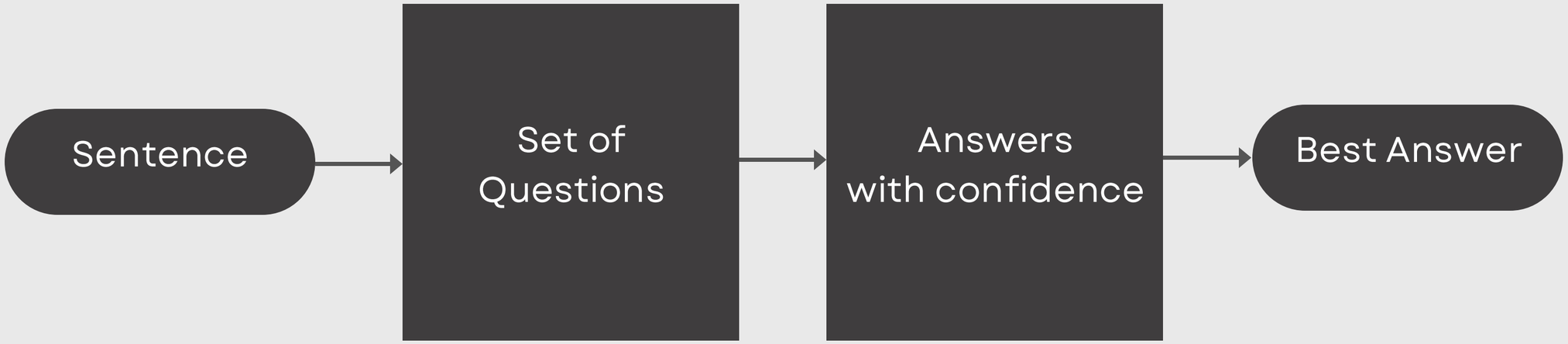
Using spacy

Label	D
PERSON	P
ORG	C
GPE	C
LOC	N ra
DATE	A
TIME	T
EVENT	N

Approach - 2 : Using QNA Method

There are several models designed to answer questions based on a given context. One such model is **distilbert-base-cased-distilled-squad**, a lightweight version of BERT fine-tuned on the SQuAD dataset for question answering tasks.

In this approach, we use a set of predefined questions to extract the required information – title, date/time, and location. For each type of information, we provide a set of relevant questions, and the model returns the most appropriate answer based on the given context.



Example


You're invited to the Halloween Party in Grand Hyatt Mumbai.
What is the event title?
What is the name of the event?
What event is being described?
What is the title of this event?
Halloween Party
0.867
Halloween Party
0.924
Halloween Party
0.910
Halloween Party
0.861
Answer : Halloween Party

THANK YOU