EXPERT ANSWERS IN A FLASH

Improving domain-specific Question-Answering

team-49/**inter-iit-techmeet**

team-49/
inter-iit-techmeet

Overview

1. Our pipeline takes a question and a set of paragraphs, and **ranks** the paragraphs based on relevance with respect to the question.

Overview

1. Our pipeline takes a question and a set of paragraphs, and **ranks** the paragraphs based on relevance with respect to the question.

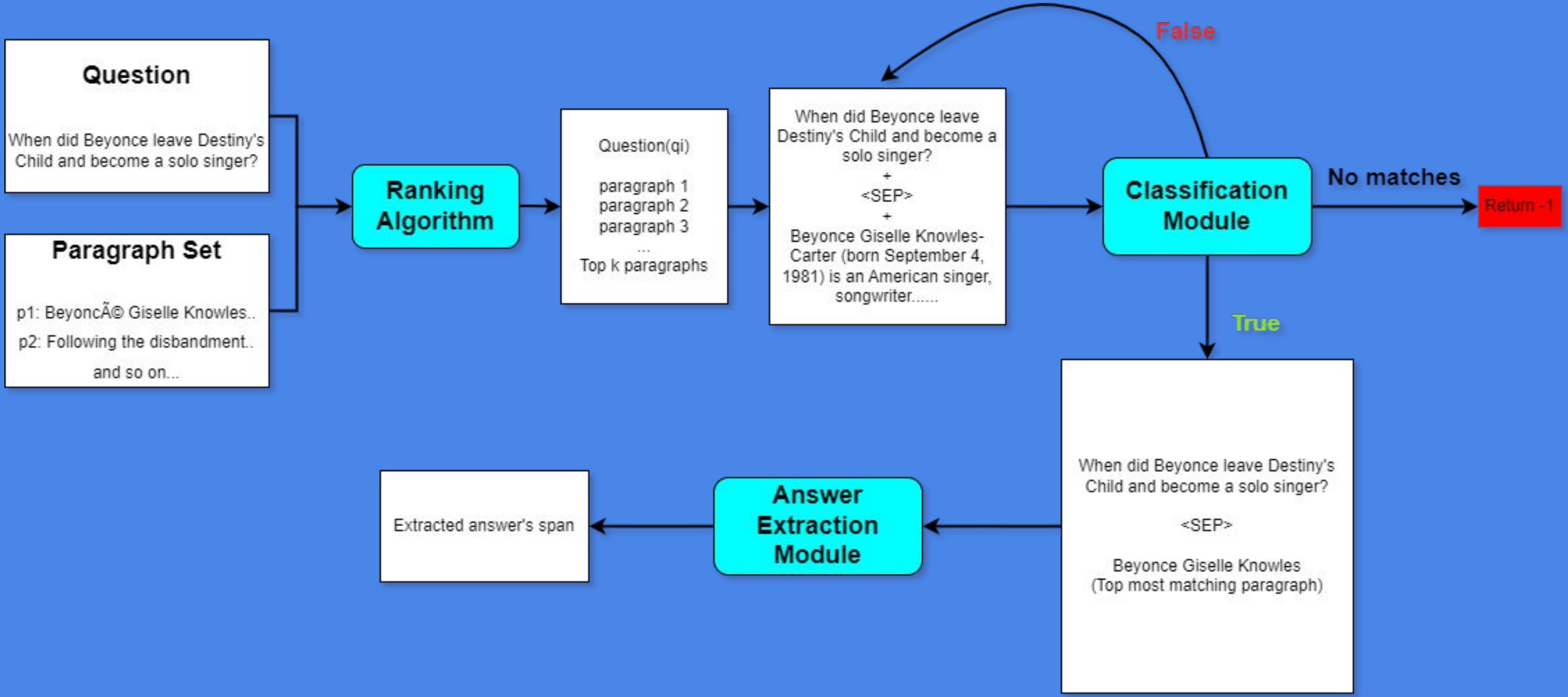
2. A Neural network **decides**,
   a. If the question can be answered by the top ranked paragraphs,
   b. Otherwise we discard the paragraph.

Overview

1. Our pipeline takes a question and a set of paragraphs, and **ranks** the paragraphs based on relevance with respect to the question.

2. A Neural network **decides**,
   a. If the question can be answered by the top ranked paragraphs,
   b. Otherwise we discard the paragraph.

3. A second neural network **extracts** the location of the answer in the paragraph.


Question
When did Beyonce leave Destiny's Child and become a solo singer?
Paragraph Set
p1: BeyoncÃ© Giselle Knowles..
p2: Following the disbandment..
and so on...
Ranking Algorithm
Question(qi)
paragraph 1
paragraph 2
paragraph 3
...
Top k paragraphs
When did Beyonce leave Destiny's Child and become a solo singer?
+
<SEP>
+
Beyonce Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter......
False
Classification Module
No matches
Return -1
True
When did Beyonce leave Destiny's Child and become a solo singer?
<SEP>
Beyonce Giselle Knowles
(Top most matching paragraph)
Answer Extraction Module
Extracted answer's span

Ranking

- **Word2Vec + KNN**
  - Word2Vec embeddings of Question and Paragraphs.
  - Finding nearest neighbours among paragraphs w.r.t questions' embedding.

Ranking

- **Word2Vec + KNN**
  - Word2Vec embeddings of Question and Paragraphs.
  - Finding nearest neighbours among paragraphs w.r.t questions’ embedding.
  - **Why it did not work?**

Ranking

- **Word2Vec + KNN**
  - Word2Vec embeddings of Question and Paragraphs.
  - Finding nearest neighbours among paragraphs w.r.t questions' embedding.
  - **Why it did not work?**
    - **Paragraphs belonging to the same theme end up with similar embeddings.**

Ranking

- **Word2Vec + KNN**
  - Word2Vec embeddings of Question and Paragraphs.
  - Finding nearest neighbours among paragraphs w.r.t questions' embedding.
  - **Why it did not work?**
    - **Paragraphs belonging to the same theme end up with similar embeddings.**
- ***Okapi* BM25**
  - BM25 (a bag-of-words based algorithm) performed better.

Synthetic data generation

- We took the **`Answer_start`**, **removed** the sentence corresponding to that index and **updated** the **`Answer_possible`** parameter to False.
    - Catered to the True/False **imbalance** in the dataset.
    - Total dataset increased from 75,056 to 1,02,166

Synthetic data generation

- We took the Answer_start, **removed** the sentence corresponding to that index and **updated** the Answer_possible parameter to False.
  - Catered to the True/False **imbalance** in the dataset.
  - Total dataset increased from 75,056 to 1,02,166
- To expand the train set size for task 2
  - **Jumbled** the sentences, preserving the answer indices to create new data points.
  - Considered **paraphrasing**, but rejected it due to time constraints (~1.5min per paragraph)

Classification

We used **Pre-trained Language + 1 Layer classification MLP head** to classify whether the question can be answered from the paragraph.

Model	Training_Accuracy	Validation_Accuracy	Test_Accuracy	Parameters
BERT	93.201	83.34	83.35	110M
ALBERT	96.332	87.34	87.34	12M
ROBERTA	95.689	89.8	89.8	125M
DISTILBERT	82.015	79.38	79.38	65M
MiniLM	93.116	89.76	89.77	33M

Extraction

1. We use **Pre-trained Language + 1 Layer MLP layer** to extract the answer start and end indices from the paragraph.

Fine-tuning

D-Day

1. Randomly paired questions and paragraphs to **create** false cases for Task 1.

Fine-tuning

D-Day

1. Randomly paired questions and paragraphs to **create** false cases for Task 1.
2. Did a simple substring matching to **locate** the answer in the paragraph.
    - Removed those instances which have more than one matching instances.

1 General
Vs
30 Specific Model

1. Why **not** 30 theme specific models?

    - Specialized models in our problem add very little value as they do **not** bring any new **domain-specific knowledge** for a theme.

1 General Vs 30 Specific Model

1. Why **not** 30 theme specific models?

    - Specialized models in our problem add very little value as they do **not** bring any new **domain-specific knowledge** for a theme.
    - Theme specific models **lack generalizability** due to limited examples for some themes

1 General
Vs
30 Specific Model

1. Why **not** 30 theme specific models?

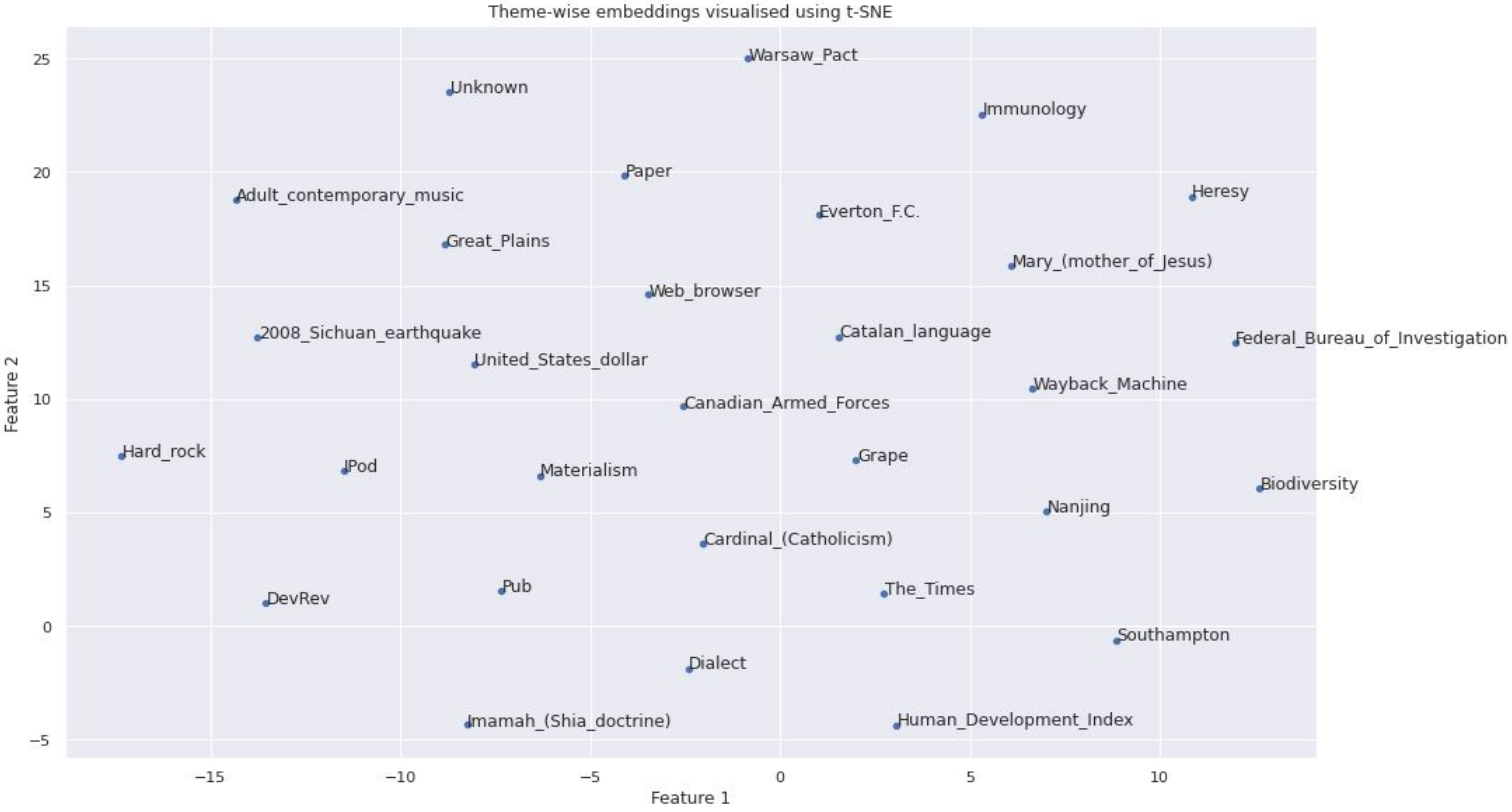
    - HDBSCAN algorithm on embeddings of Concatenated paragraphs on the same themes resulted in **only two clusters**.

1 General
Vs
30 Specific Model

1. Why **not** 30 theme specific models?

    - HDBSCAN algorithm on embeddings of Concatenated paragraphs on the same themes resulted in **only two clusters**.

    - t-SNE visualization of theme-wise embeddings showed **30 separate points**.


Theme-wise embeddings visualised using t-SNE
Warsaw_Pact
Unknown
Immunology
Paper
Adult_contemporary_music
Heresy
Everton_F.C.
Great_Plains
Mary_(mother_of_Jesus)
Web_browser
2008_Sichuan_earthquake
Catalan_language
Federal_Bureau_of_Investigation
United_States_dollar
Wayback_Machine
Canadian_Armed_Forces
Hard_rock
IPod
Materialism
Grape
Biodiversity
Nanjing
Cardinal_(Catholicism)
Pub
DevRev
The_Times
Southampton
Dialect
Imamah_(Shia_doctrine)
Human_Development_Index
Feature 1
Feature 2
−15
−10
−5
0
5
10
15
20
25

Top-3 classes with highest average prediction time per question

Class	Average Prediction Time per Question (sec)
Materialism	4.12
Federal_Bureau_of_Investigation	3.41
Human_Development_Index	3.40

Top-3 classes with lowest average prediction time per question

Class	Average Prediction Time per Question (sec)
The_Times	1.23
Nanjing	1.30
DevRev	1.32

Runtime Analysis

Pipeline Sections	Training time (min)
Synthetic Data Generation	15
Task1_MiniLM	360
Synthetic Data Generation	45
Task2_MiniLM	75
Fine tuning data generation	15
Inference	150

Thank you