00
01
02
03

FPT Education
FPT UNIVERSITY

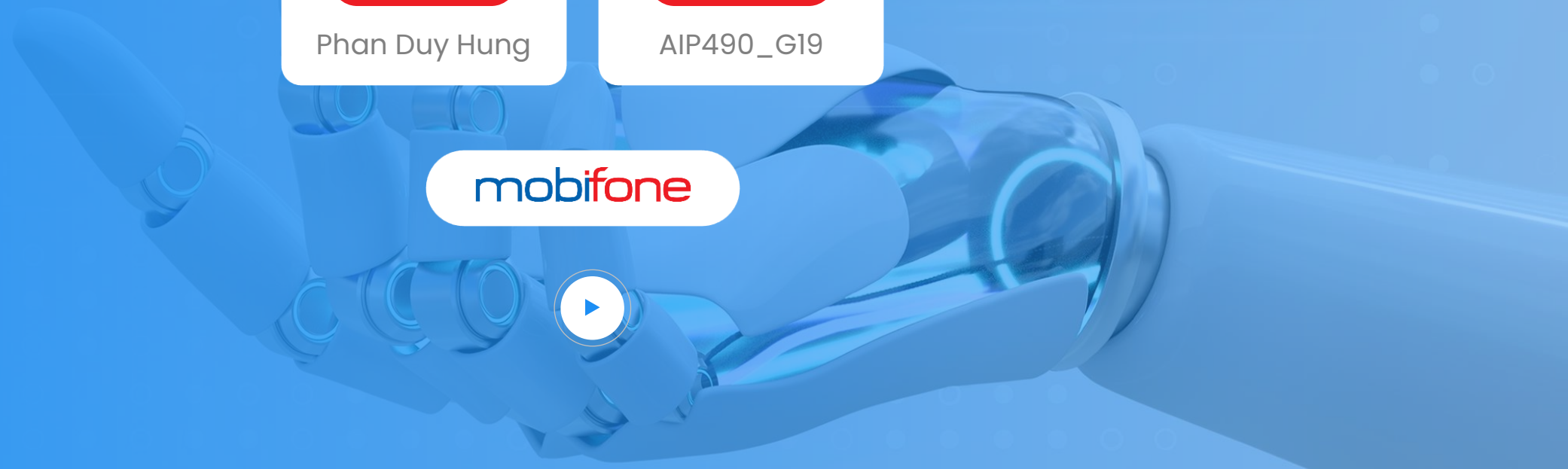
Enhancing Semantic Search through Domain

Adaptation: A Case Study on the Mobifone News Dataset

Mentor
Phan Duy Hung

Group
AIP490_G19

mobifone



OUR **MEMBERS.**

PHAN DUY HUNG

Supervisor

Nguyen Trieu Ngoc Huyen

HE150374

Thach Duc Long

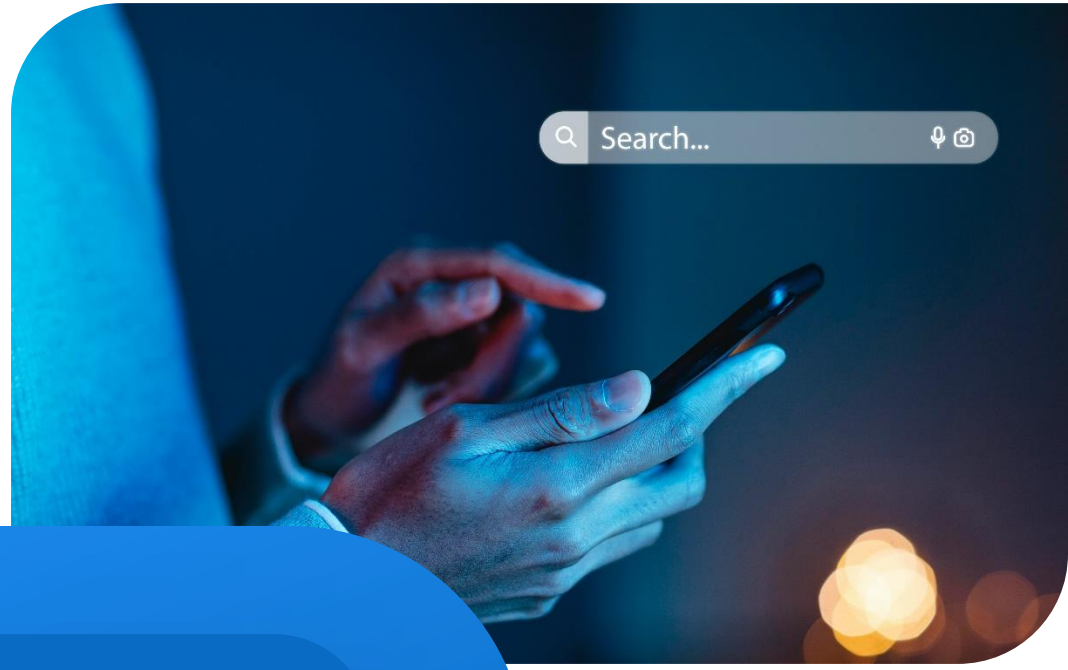
HE150206

Nguyen Minh Hieu

HE150296

FPT University

The outline



01.

Introduction

02.

Methodology

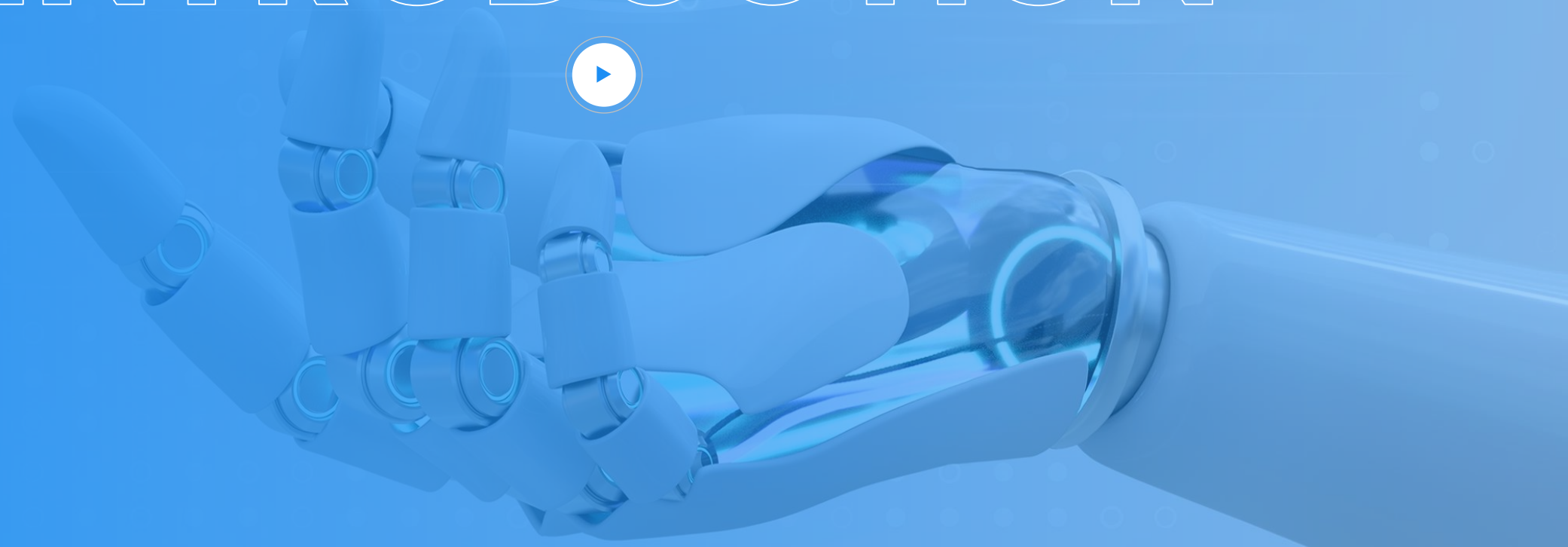
03.

Conclusion
Future Works

FPT Education

FPT UNIVERSITY

01 INTRODUCTION



Problem & Motivation

Mobifone is a Telecommunication Company in Vietnam that consumes a massive amount of news every day, both internal and external to their customers. These pieces of news come in the form of promotional SMS, official webpage announcements, internal department reports, etc. All these text data need a solution for customers, support lines, and internal access to search and retrieve in seconds.

The existing system is based on a keyword database "LIKE" search. Our mission is to upgrade this system to get a higher search intelligence whilst reducing search time.

01. Problem and motivation

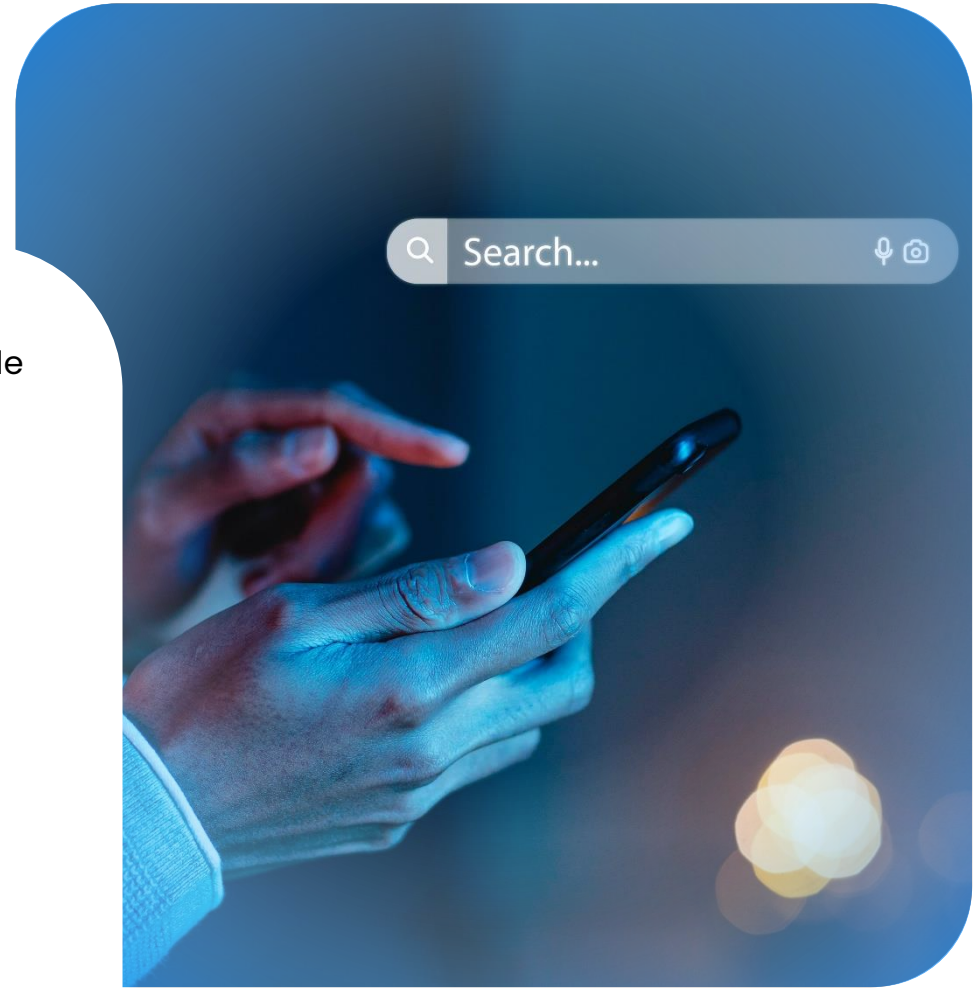
Semantic Search

- Semantic search considering the meaning behind the words to understand user intent and context. Deliver contextually relevant results by comprehending the relationships between words and the overall context of the query.

- Semantic search is an evolution in both accuracy and flexibility of the search engine. Unlike traditional keyword-based searches, it delves deeper by comprehending the semantics behind words and user queries, resulting in more accurate and relevant search outcomes.

02. **Related work**

- Some well-known datasets commonly used for general purposes include Microsoft Research Paraphrase Corpus (MRPC), SemEval
- These datasets include Hypothesis - Premise – Label
- Mobifone News dataset only contain Title – Description - Content



premise string · *lengths* 15 — 139	**hypothesis** string · *lengths* 14 — 144	**relatedness_score** float32 1 — 5	**entailment_judgment** class label 3 classes
A group of kids is playing in a yard and…	A group of boys in a yard is playing and a…	4.5	0 NEUTRAL
A group of children is playing in the house…	A group of kids is playing in a yard and…	3.2	0 NEUTRAL
The young boys are playing outdoors and…	The kids are playing outdoors near a man…	4.7	1 ENTAILMENT
The kids are playing outdoors near a man…	A group of kids is playing in a yard and…	3.4	0 NEUTRAL

SemEval Dataset Structure

sentence1 string	**sentence2** string	**similarity_score** float32
A plane is taking off.	An air plane is taking off.	5
A man is playing a large flute.	A man is playing a flute.	3.8
A man is spreading shreded cheese on a pizza.	A man is spreading shredded cheese on an uncooked pizza.	3.8
Three men are playing chess.	Two men are playing chess.	2.6
A man is playing the cello.	A man seated is playing the cello.	4.25

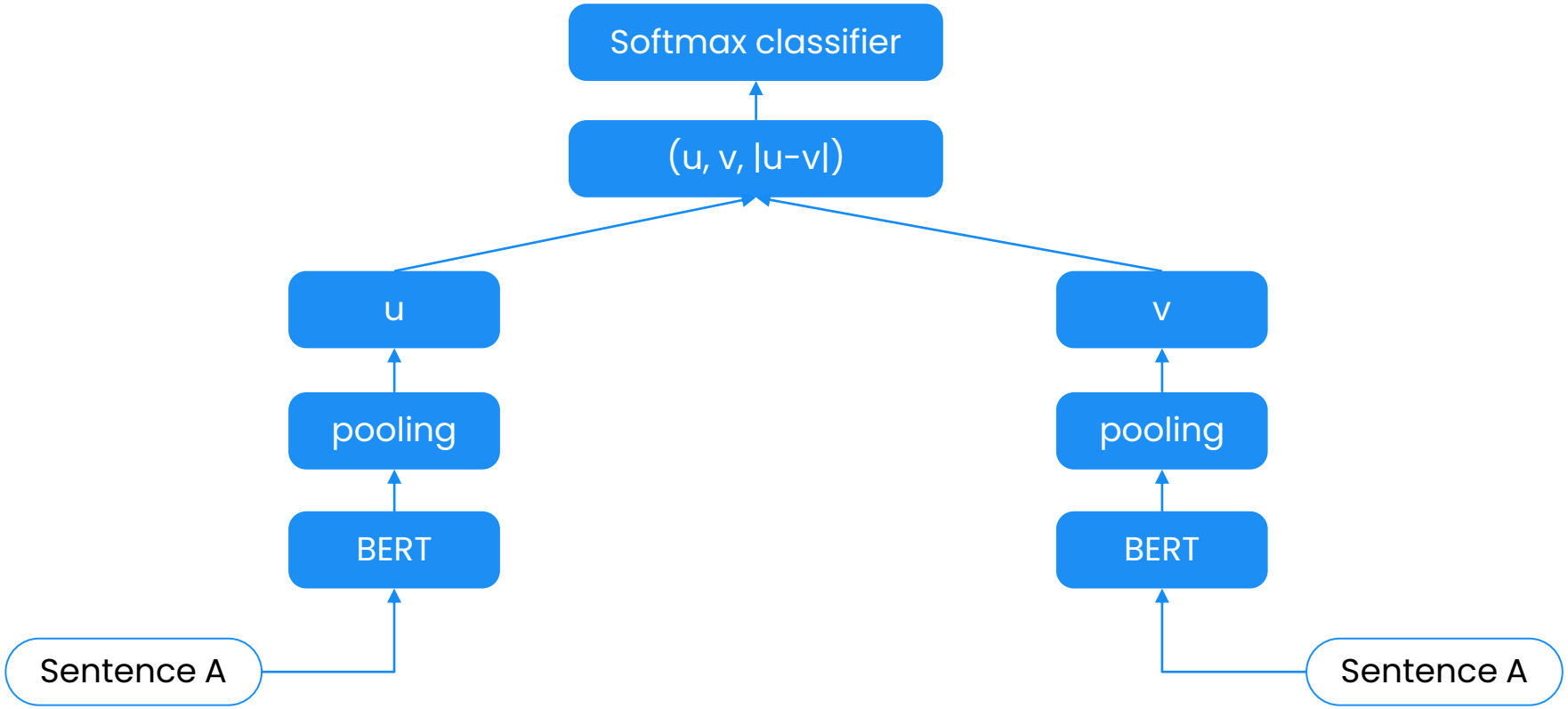
A normal Semantic Similarity Scoring Dataset Example

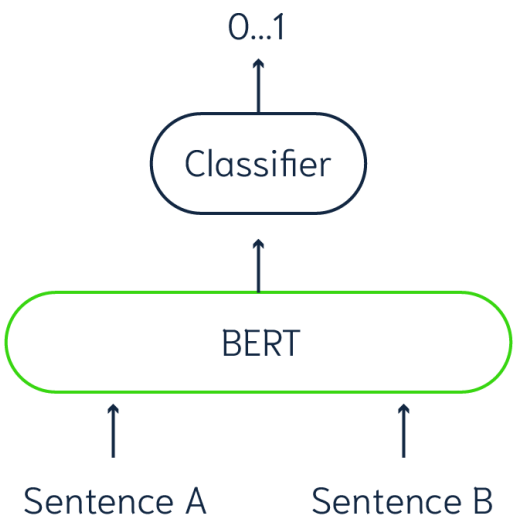
Tiêu đề	Nội dung	Mô tả
Gói cước Rockstorm (thử nghiệm)	Mức cước: khai báo theo mức cước của gói cước QStudent. Các ưu đãi: Ngoài ưu đãi của gói cước Q-Student , KH còn được thêm các ưu đãi sau: Miễn phí gói data Zing: KH được tặng gói cước data Zing trong 01 chu kỳ đầu tiên khi kích hoạt thuê bao. Các chu kỳ tiếp theo nếu KH gia hạn sẽ tính cước theo mức quy định của gói. Được mặc định tham gia cộng đồng RockStorm và hưởng đầy đủ các quyền lợi, ưu đãi dành cho thành viên trong cộng đồng (xem tại Tab Ưu đãi cho cộng đồng RockStorm). NỘI DUNG CƯỚC (Đã có VAT) Gọi nội mạng khoảng 1.18 0đ/phút 118đ/6s + 19,67đ/1s Gọi nhóm (kể cả khi roaming với Vina) khoảng 708đ/phút 70,8đ/6s + 11,8đ/1s Gọi TB trong cộng đồng Rockstorm khoảng 708đ/phút 70,8đ/6s + 11,8đ/1s Gọi liên mạng khoảng 1. 380đ/phút 138đ/6s + 23đ/1s S MS nội mạng 99đ/SMS SMS liên mạng 250đ/SMS SMS cố định VNPT 290đ/SMS 1. Đối tượng áp dụng: Đối tượng áp dụng: Thuê bao đang hoạt động trên mạng MobiFone và các thuê bao đăng ký gói cước RockStorm. Đăng ký tham gia: Thuê bao phát triển mới sử dụng gói cước RockStorm: KH được mặc định tham gia cộng đồng ngay khi kích hoạt mà không cần đăng ký. Thuê bao MobiFone đang hoạt động trên mạng: KH có thể đăng ký tham gia cộng đồng RockStorm bằng cách soạn tin nhắn: DK_RS gửi 999 (trong đó " "_" " là khoảng trống khi soạn tin nhắn). Sau khi tham gia cộng đồng RockStorm, KH sẽ được ngay hưởng toàn bộ các ưu đãi dành cho thành viên trong cộng đồng. 2. Ưu đãi cho thành viên trong cộng đồng: a. Giảm cước gọi giữa các thành viên trong cộng đồng RockStorm: Các thuê bao trong cộng đồng RockStorm sẽ được hưởng mức cước gọi ưu đãi khi gọi cho nhau: 708 đồng/phút (đã bao gồm thuế GTGT), không phân biệt gói cước mà KH đang sử dụng. Trong đó: Block 06 giây đầu: 70,8 đồng/06 giây Block 01 giây tiếp theo: 11,8 đồng/01 giây Mức cước này chỉ áp dụng khi các thành viên gọi cho nhau. Khi thành viên gọi đến các thuê bao không nằm trong cộng đồng RockStorm: áp dụng theo mức cước của gói cước mà KH đang sử dụng. b. Cập nhật thông tin miễn phí về chương trình RockStorm: Trong thời gian diễn ra chương trình RockStorm, thành viên cộng đồng RockStorm sẽ được cập nhật miễn phí gói thông tin hàng ngày về chương trình. Thuê bao phải đăng ký để được nhận gói thông tin miễn phí. c. Giảm cước tải nhạc chuông chờ RockStorm: Ngay khi đăng ký thành viên cộng đồng RockStorm, thành viên sẽ được tặng miễn phí bộ nhạc chuông chờ RockStorm. Giảm 50% cước tải nhạc chuông chờ RockStorm trong các lần tải tiếp theo. d. Tặng quà cho các thuê bao trong cộng đồng có mức sử dụng cao nhất hàng tháng: Thời gian để tính và xếp hạng thành viên: 01 chu kỳ được tính từ ngày 01 đến ngày cuối cùng của tháng. Thời điểm chốt dữ liệu để xếp hạng: ngày 05 hàng tháng. Số tiền sử dụng để xếp hạng: Đối với thuê bao trả trước: là tổng số tiền tiêu dùng trong tài khoản chính trong chu kỳ trước. Đối với thuê bao trả sau: là số tiền KH thực phải nộp trong chu kỳ trước, đã bao gồm thuế GTGT (không bao gồm các khoản nợ trước, chiết khấu thương mại, khuyến mại,...). Quy định về xếp hạng thành viên: vào ngày 05 hàng tháng, hệ thống sẽ chốt danh sách thành viên trong cộng đồng RockStorm và tiến hành xếp hạng thành viên theo mức sử dụng của chu kỳ (tháng) liền kề trước đó. 10 thuê bao có mức sử dụng cao nhất trong tháng sẽ được nhận quà tặng của MobiFone (KH sẽ nhận được tin nhắn thông báo về việc được tặng quà). Tặng quà cho thuê bao có mức sử dụng cao: 10 thuê bao có mức sử dụng cao nhất trong tháng sẽ được nhận quà tặng là bộ nhạc chuông chờ mới nhất của MobiFone hoặc các ấn vật phẩm, ấn phẩm liên quan đến chương trình RockStorm. e. Tham dự các sự kiện, chương trình liên quan đến RockStorm: Giao lưu, họp báo giới thiệu show diễn, tặng vé xem RockStorm miễn phí cho các năm tiếp theo,... TT Tình huống Nội dung tin nhắn 1 Thuê bao đăng ký tham gia cộng đồng RockStorm 1.1 Thuê bao MobiFone chưa tham gia cộng đồng Chao mung ban tro thanh thanh vien cong dong MobiFone RockStorm. Ban se duoc dong hanh voi chuong trinh RockStorm va nhan nhung uu dai dac biet tu MobiFone.	Trong thời gian diễn ra chương trình RockStorm, thành viên cộng đồng RockStorm sẽ được cập nhật miễn phí gói thông tin hàng ngày về chương trình. Thuê bao phải đăng ký để được nhận gói thông tin miễn phí. c. Giảm cước tải nhạc chuông chờ RockStorm: Ngay khi đăng ký thành viên cộng đồng RockStorm, thành viên sẽ được tặng miễn phí bộ nhạc chuông chờ RockStorm. Giảm 50% cước tải nhạc chuông chờ RockStorm trong các lần tải tiếp theo.

MobiFone News Dataset contains only Title, Content and Description

02. **Related work**

Traditional BERT uses a cross-encoder for semantic comparison that accepts two sentences as input and outputs a target value



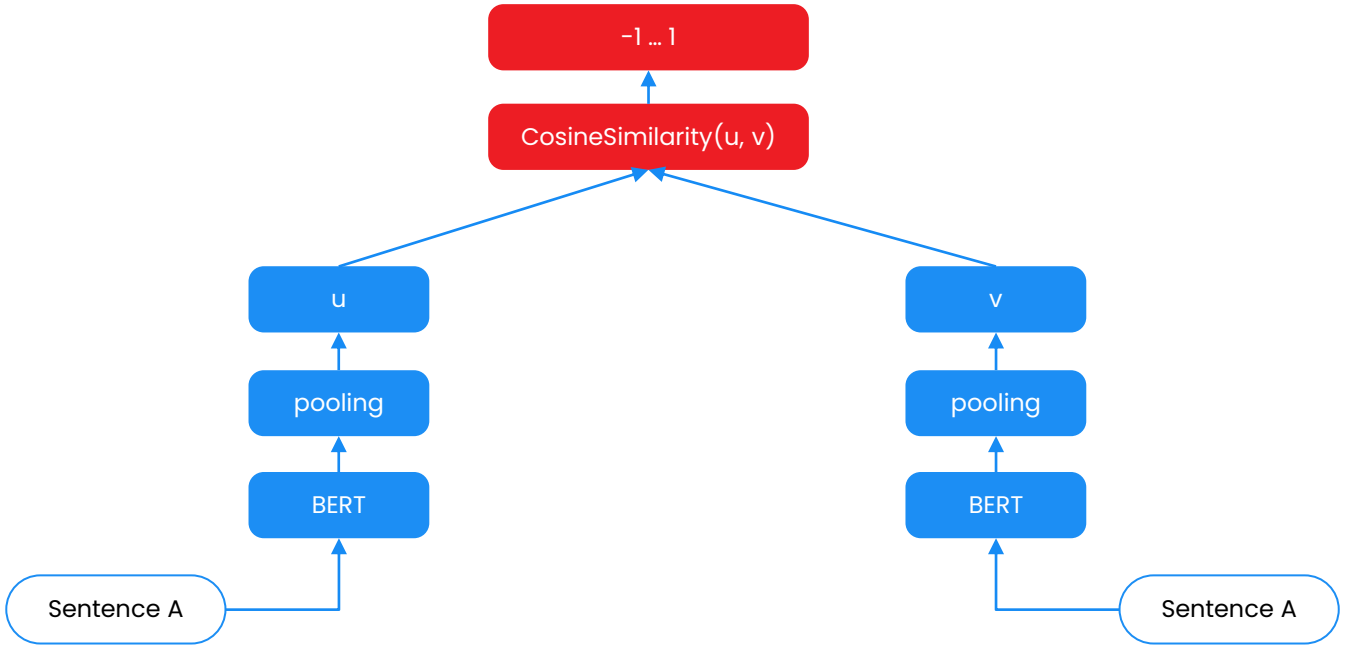


When using cross-encoder, we must provide two input vector simultaneously to get their similarity. In contrast, a bi-encoder accept each input independently, return their embedding vectors which can then be used with a similarity metric.

- Due to the massive computations, a cross-encoder is unsuitable in the search engine use case.
- A search on a knowledge base of n=10,000 documents will take 10,000 model inferences.
- This number may multiply how the preprocess operation splits each raw document due to the BERT-model maximum accepted sequence length

Reimers and Gurevych introduce a new architecture using two identical BERT as featurizer and train simultaneously two BERT with a similarity metric so that BERT encode two semantically similar sentence into high metrically similar vectors

Using a pre-encoded knowledge base with an optimized index structure can reduce the query time of a search from days to a few milliseconds.

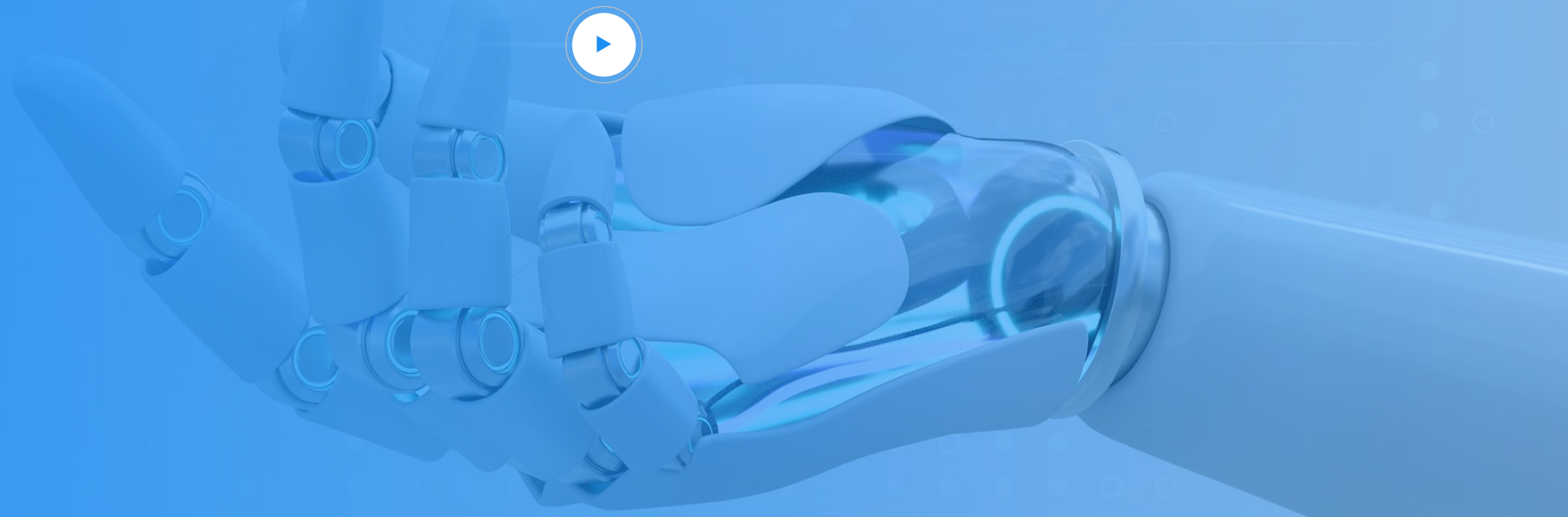


FPT Education

FPT UNIVERSITY

00
01
02
03

02 METHODOLOGY



01. Data Preparation

Raw filtering of HTML entities into corresponding Unicode characters

```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```

```
<Mobile> Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ kích
hoạt & nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
期间中通过 Fastpay 方式充值将获得赠送 50%的充值面值 。不限制充值次数
:D
là trò chơi trực tuyến giúp người chơi là các thuê bao MobiFone tham gia
dự đoán tỷ số các trận đấu tại các giải đấu bóng đá và nhận các giải
thưởng. Các giải đấu cung cấp dịch vụ dự đoán:
```

01. Data Preparation

Remove HTML tags with whitespace and Convert characters from Windows-1252 encoding to UTF-8


```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```



```
 Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ kích hoạt & nạp
tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
期间中通过 Fastpay 方式充值将获得赠送 50%的充值面值 。不限制充值次数
:D
là trò chơi trực tuyến giúp người chơi là các thuê bao MobiFone tham gia
dự đoán tỷ số các trận đấu tại các giải đấu bóng đá và nhận các giải
thưởng. Các giải đấu cung cấp dịch vụ dự đoán:
```

01. Data Preparation

Cleans and removes Email, URL, Mention, number recognition and grouping of numbers


```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```



```
  Đối tượng: các bộ hòa mạng là MobiCard, Mobi 4 U và MobiQ kích hoạt &
nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi 5 feedback 64 &
SĐt: 0967969264 or 0 967969264 or x 009983
mail :
Ngày: 12/09/2023 :)
期间中通过 Fastpay 方式充值将获得赠送 50 %的充值面值 。不限制充值次数
:D
là trò chơi trực tuyến giúp người chơi là các thuê bao MobiFone tham gia
dự đoán tỷ số các trận đấu tại các giải đấu bóng đá và nhận các giải
thưởng. Các giải đấu cung cấp dịch vụ dự đoán:
```

01. Data Preparation

Remove emoji and convert data to lowercase


```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```



```
  đối tượng: các bộ hòa mạng là mobicard, mobi 4 u và mobiq kích hoạt &
nạp tiền.
chấtlượng tốt, this is sản phẩm! trải nghiệm rat tuyetvoi,
hãy cho chung toi 5 feedback 64 &
sđt: 0967969264 or 0 967969264 or x 009983
mail :
ngày: 12/09/2023
期间中通过 fastpay 方式充值将获得赠送 50 %的充值面值 。不限制充值次数

là trò chơi trực tuyến giúp người chơi là các thuê bao mobifone tham gia
dự đoán tỷ số các trận đấu tại các giải đấu bóng đá và nhận các giải
thưởng. các giải đấu cung cấp dịch vụ dự đoán:
```

01. Data Preparation

Remove non-numeric characters, spaces, latin letters, and underscores
Replace consecutive characters with a single character (including tabs and newlines)
Ex: dashes, spaces


```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```



```
đối tượng: các bộ hòa mạng là mobicard, mobi 4 u và mobiq kích hoạt nạp
tiền. chấtlượng tốt, this is sản phẩm trải nghiệm rat tuyetvoi, hãy cho
chung toi 5 feedback 64 sđt: 0967969264 or 0 967969264 or x 009983 mail :
ngày: 12/09/2023 fastpay 50 là trò chơi trực tuyến giúp người chơi là các
thuê bao mobifone tham gia dự đoán tỷ số các trận đấu tại các giải đấu
bóng đá và nhận các giải thưởng. các giải đấu cung cấp dịch vụ dự đoán:
```

01. Data Preparation

Standardize Vietnamese word marks


```
&lt;Mobile&gt; Đối tượng: các bộ hòa mạng là MobiCard, Mobi4U và MobiQ
kích hoạt &amp; nạp tiền.
ChấtLượng Tốt, This is sản phẩm! trải nghiệm rat tuyetvoi,
Hãy cho chung toi5feedback64 &amp;
SĐt: 0967969264 or 0 967969264 or x009983
mail : Longnmhe150296@fpt.edu.vn
Ngày: 12/09/2023 :)
&#26399;&#38388;&#20013;&#36890;&#36807;Fastpay&#26041;&#24335;&#20805;&#
20540;&#23558;&#33719;&#24471;&#36192;&#36865;
50%&#30340;&#20805;&#20540;&#38754;&#20540;
&#12290;&#19981;&#38480;&#21046;&#20805;&#20540;&#27425;&#25968;
:D
&#108;&#224;&#32;&#116;&#114;&#242;&#32;&#99;&#104;&#417;&#105;&#32;&#116
;&#114;&#7921;&#99;&#32;&#116;&#117;&#121;&#7871;&#110;&#32;&#103;&#105;&
#250;&#112;&#32;&#110;&#103;&#432;&#7901;&#105;&#32;&#99;&#104;&#417;&#10
5;&#32;&#108;&#224;&#32;&#99;&#225;&#99;&#32;&#116;&#104;&#117;&#234;&#32
;&#98;&#97;&#111;&#32;&#77;&#111;&#98;&#105;&#70;&#111;&#110;&#101;&#32;&
#116;&#104;&#97;&#109;&#32;&#103;&#105;&#97;&#32;&#100;&#7921;&#32;&#273;
&#111;&#225;&#110;&#32;&#116;&#7927;&#32;&#115;&#7889;&#32;&#99;&#225;&#9
9;&#32;&#116;&#114;&#7853;&#110;&#32;&#273;&#7845;&#117;&#32;&#116;&#7841
;&#105;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#78
45;&#117;&#32;&#98;&#243;&#110;&#103;&#32;&#273;&#225;&#32;&#118;&#224;&#
32;&#110;&#104;&#7853;&#110;&#32;&#99;&#225;&#99;&#32;&#103;&#105;&#7843;
&#105;&#32;&#116;&#104;&#432;&#7903;&#110;&#103;&#46;&#32;&#67;&#225;&#99
;&#32;&#103;&#105;&#7843;&#105;&#32;&#273;&#7845;&#117;&#32;&#99;&#117;&#
110;&#103;&#32;&#99;&#7845;&#112;&#32;&#100;&#7883;&#99;&#104;&#32;&#118;
&#7909;&#32;&#100;&#7921;&#32;&#273;&#111;&#225;&#110;&#58;
```



```
đối tượng: các bộ hòa mạng là mobicard, mobi 4 u và mobiq kích hoạt nạp
tiền. chấtlượng tốt, this is sản phẩm trải nghiệm rat tuyetvoi, hãy cho
chung toi 5 feedback 64 sđt: 0967969264 or 0 967969264 or x 009983 mail :
ngày: 12/09/2023 fastpay 50 là trò chơi trực tuyến giúp người chơi là các
thuê bao mobifone tham gia dự đoán tỷ số các trận đấu tại các giải đấu
bóng đá và nhận các giải thưởng. các giải đấu cung cấp dịch vụ dự đoán:
```

Title cleaning

(MBF KV 3) Chặn cắt số vi phạm quảng cáo rao vặt tại Đà Nẵng	Chặn cắt số vi phạm quảng cáo rao vặt tại Đà Nẵng
Buzz Me (Báo cuộc gọi nhỡ tự động) 9283	Buzz Me Báo cuộc gọi nhỡ tự động 9283

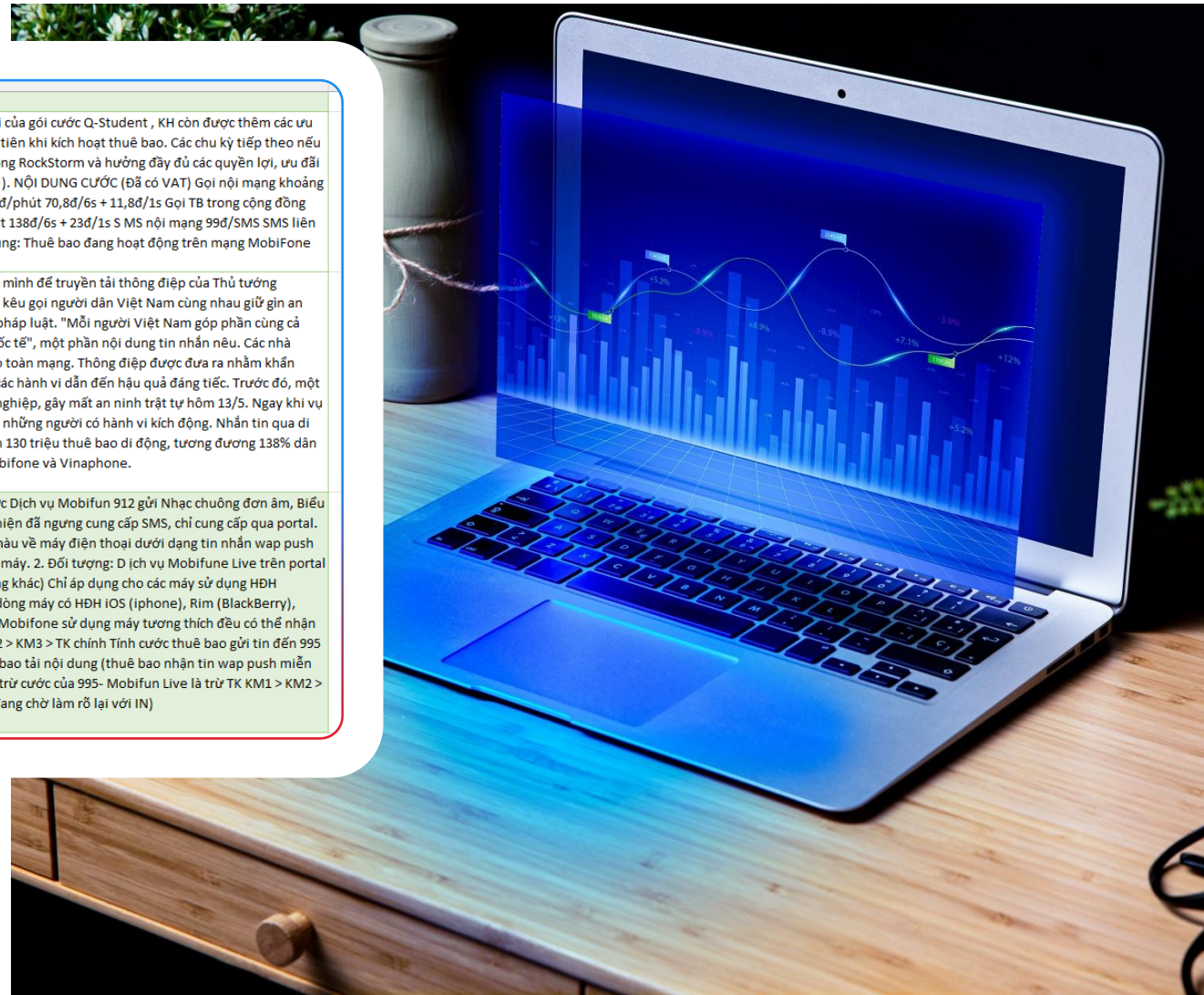
News Content Processing

Run a Language Detection Pretrained model to clear all leftover data which is not Vietnamese

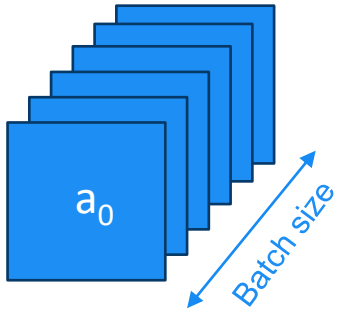
Cut News with content longer than 1000 words down to first 10 sentences. (Using <EOS> token of PhoBERT tokenizer)

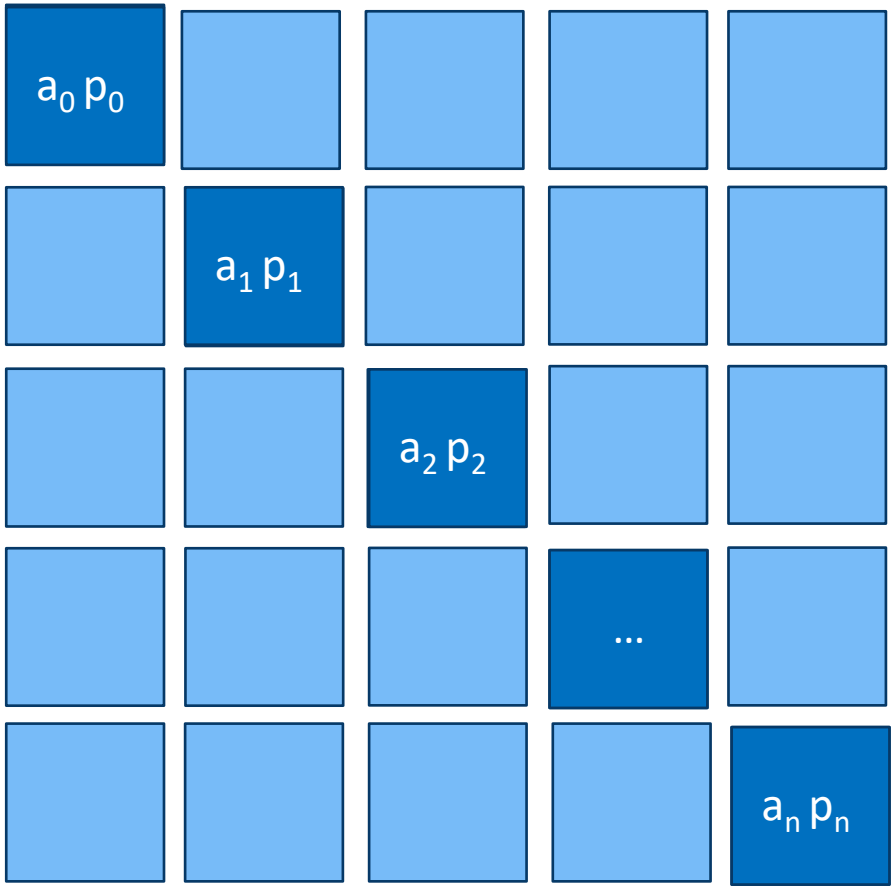
Data After Cleaning

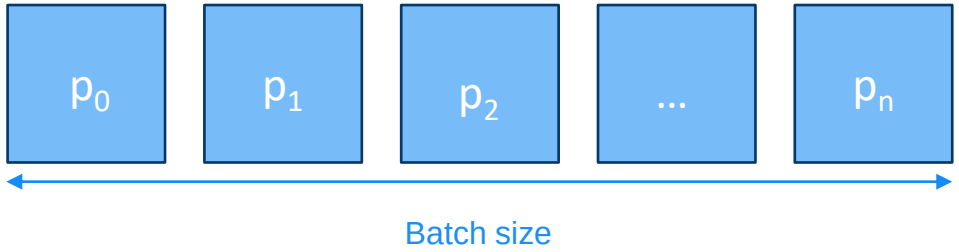
Title	Content
Gói cước Rockstorm thử nghiệm	Mức cước: khai báo theo mức cước của gói cước QStudent. Các ưu đãi: Ngoài ưu đãi của gói cước Q-Student , KH còn được thêm các ưu đãi sau: Miễn phí gói data Zing: KH được tặng gói cước data Zing trong 01 chu kỳ đầu tiên khi kích hoạt thuê bao. Các chu kỳ tiếp theo nếu KH gia hạn sẽ tính cước theo mức quy định của gói. Được mặc định tham gia cộng đồng RockStorm và hưởng đầy đủ các quyền lợi, ưu đãi dành cho thành viên trong cộng đồng (xem tại Tab Ưu đãi cho cộng đồng RockStorm). NỘI DUNG CƯỚC (Đã có VAT) Gọi nội mạng khoảng 1.18 0đ/phút 118đ/6s + 19,67đ/1s Gọi nhóm (kể cả khi roaming với Vina) khoảng 708đ/phút 70,8đ/6s + 11,8đ/1s Gọi TB trong cộng đồng Rockstorm khoảng 708đ/phút 70,8đ/6s + 11,8đ/1s Gọi liên mạng khoảng 1. 380đ/phút 138đ/6s + 23đ/1s S MS nội mạng 99đ/SMS SMS liên mạng 250đ/SMS SMS cố định VNPT 290đ/SMS 1. Đối tượng áp dụng: Đối tượng áp dụng: Thuê bao đang hoạt động trên mạng MobiFone và các thuê bao đăng ký gói cước RockStorm.
Nhà mạng nhắn tin truyền thông điệp của Thủ tướng	Từ chiều ngày 15/5, các nhà mạng tại Việt Nam đồng loạt nhắn tin đến thuê bao của mình để truyền tải thông điệp của Thủ tướng Nguyễn Tấn Dũng về việc bảo đảm an ninh trật tự trong bối cảnh hiện nay. Nội dung kêu gọi người dân Việt Nam cùng nhau giữ gìn an ninh trật tự, đoàn kết, không nghe theo kích động của kẻ xấu để hành động trái với pháp luật. "Mỗi người Việt Nam góp phần cùng cả nước bảo vệ chủ quyền thiêng liêng của Tổ quốc theo đúng luật pháp nước ta và quốc tế", một phần nội dung tin nhắn nêu. Các nhà mạng cho biết có 3 mẫu tin nhắn với thông tin tương đồng nhau sẽ gửi đến thuê bao toàn mạng. Thông điệp được đưa ra nhằm khẩn trương đồng b ộ các biện pháp, chủ động và kiên quyết ngăn chặn, không để xảy ra các hành vi dẫn đến hậu quả đáng tiếc. Trước đó, một nhóm người xấu đã kích động các công nhân tại Bình Dương phá hoại tài sản doanh nghiệp, gây mất an ninh trật tự hôm 13/5. Ngay khi vụ việc xảy ra, các cơ quan chức năng Việt Nam đã khẩn trương trấn áp và xử lý nghiêm những người có hành vi kích động. Nhắn tin qua di động là kênh tuyên truyền thông tin hiệu quả tới người dân bởi tại Việt Nam có hơn 130 triệu thuê bao di động, tương đương 138% dân số. Trong đó Viettel có gần 60 triệu khách hàng, tương đương tổng thuê bao của Mobifone và Vinaphone.
MobiFun Live.	1. Giới thiệu: Web Portal cung cấp hình thức tải nhạc, hình, logo bao gồm 2 hình thức Dịch vụ Mobifun 912 gửi Nhạc chuông đơn âm, Biểu tượng (logo đen trắng- tức logo mạng) trực tiếp về ĐT, không yêu cầu kết nối data, hiện đã ngưng cung cấp SMS, chỉ cung cấp qua portal. Dịch vụ Mobifun Live gửi Nhạc đa âm sắc (midi), có lời (mp3), hình màu (jpg), logo màu về máy điện thoại dưới dạng tin nhắn wap push chứa đường link. Người nhận kết nối đường link thông qua GPRS để tải nội dung về máy. 2. Đối tượng: D ịch vụ Mobifune Live trên portal chỉ có thể gửi t in nhắn Wap push đến tất cả thuê bao Mobifone (không gửi đến mạng khác) Chỉ áp dụng cho các máy sử dụng HĐH Symbian, Windows Mobile có khả năng nhận tin nhắn wappush. KHÔNG hỗ trợ các dòng máy có HĐH iOS (iphone), Rim (BlackBerry), Android (Sonyericson Experia),... 3. Điều kiện sử dụng: Tất cả thuê bao thuộc mạng Mobifone sử dụng máy tương thích đều có thể nhận được tin Wap push của dịch vụ Mobifune Live. Tuy nhiên để có thể tải TK KM1 > KM2 > KM3 > TK chính Tính cước thuê bao gửi tin đến 995 Tải nội dung T ính theo gói cước Internet mà thuê bao đang sử dụng Tính cước thuê bao tải nội dung (thuê bao nhận tin wap push miễn phí nhưng khi tải nội dung sẽ bị tính cước data) Ghi chú: IN ICC quy định nguyên tắc trừ cước của 995- Mobifun Live là trừ TK KM1 > KM2 > KM3 > TK chính , tuy nhiên thực tế test chỉ trừ vào TK chính (và đang trừ 1710đ/lần- đang chờ làm rõ lại với IN)



Multiple Negative Ranking





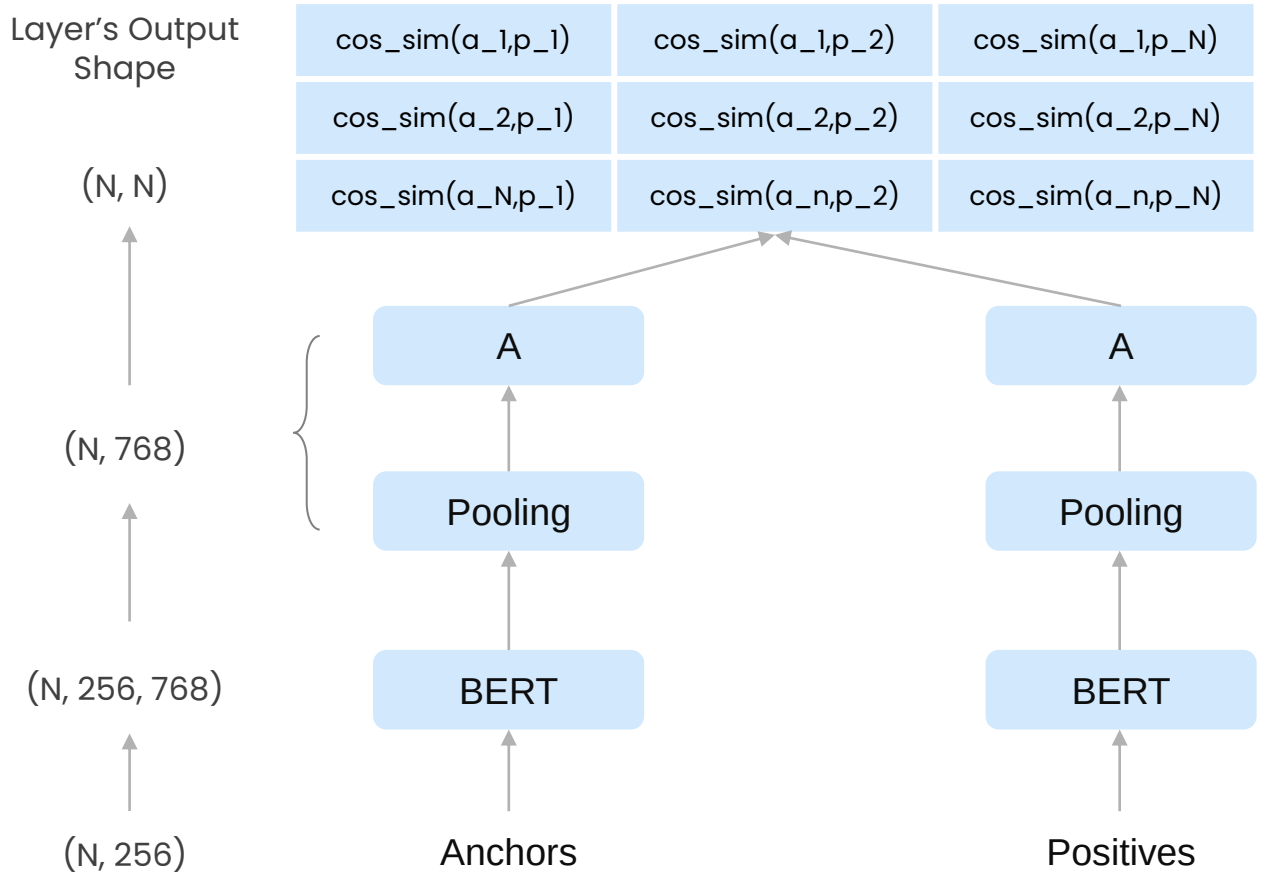


Why Multiple Negative Ranking Loss

- As our data only contains news titles and content, there will be no contradiction or neutral annotation as in ideal datasets.
- Using any classification will halve performance because there will be a class imbalance between positive and negative.
- We will batch sentences and marking all non-negative labels with a low score.

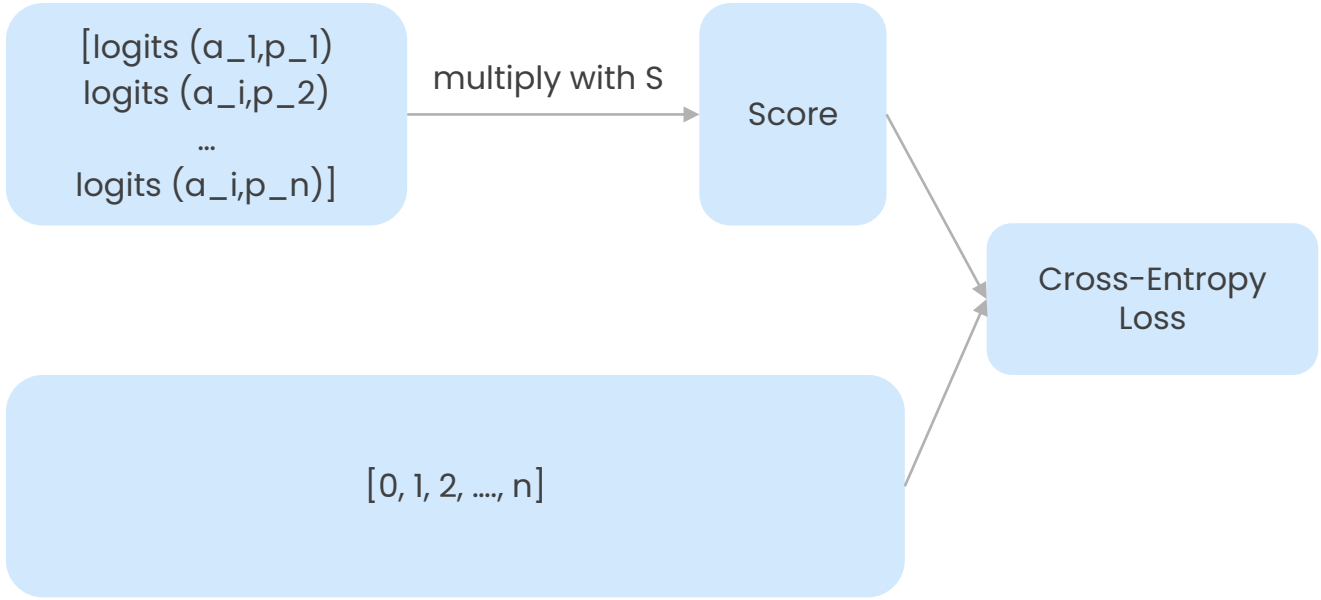
Model architecture

Our PyTorch implementation loads a single BERT network with pre-trained weights from the 'vinai/phobert-base'. This BERT network processes two batches of sentences separately in the training process, creating a "siamese"-like network. Our design guaranteed the identical properties between BERT networks of SBERT



Optimization Target

With n as batch size, the training targets to minimize cross-entropy loss. The label is a list with length=n, indicating the index of p that a_i must match.



System scenario

The search engine will be running on the following system specifications:

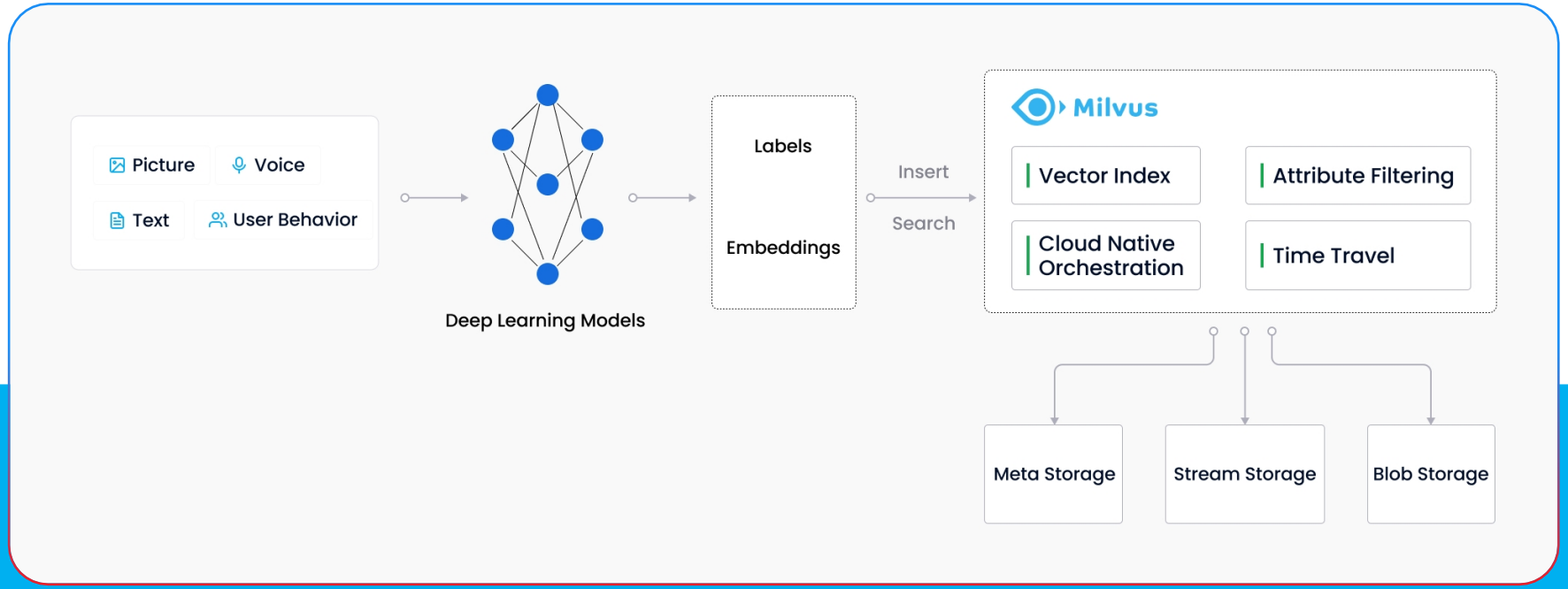
- 4 cores 3.40GHZ Intel CPU with AVX2 support
- 64GB RAM
- Generous volume amount of multiple Gen 3.0 PCI-e NVME SSDs
- NO GPU or GPU turned off

Performance threshold that the search engine must achieve:

- Below 1 second per search query at any system condition.
- Both search engine and communication API take under 32GB of RAM.

Milvus vector database

Milvus is an open-source project introduced in 2019 to store, index, and manage massive embedding vectors.

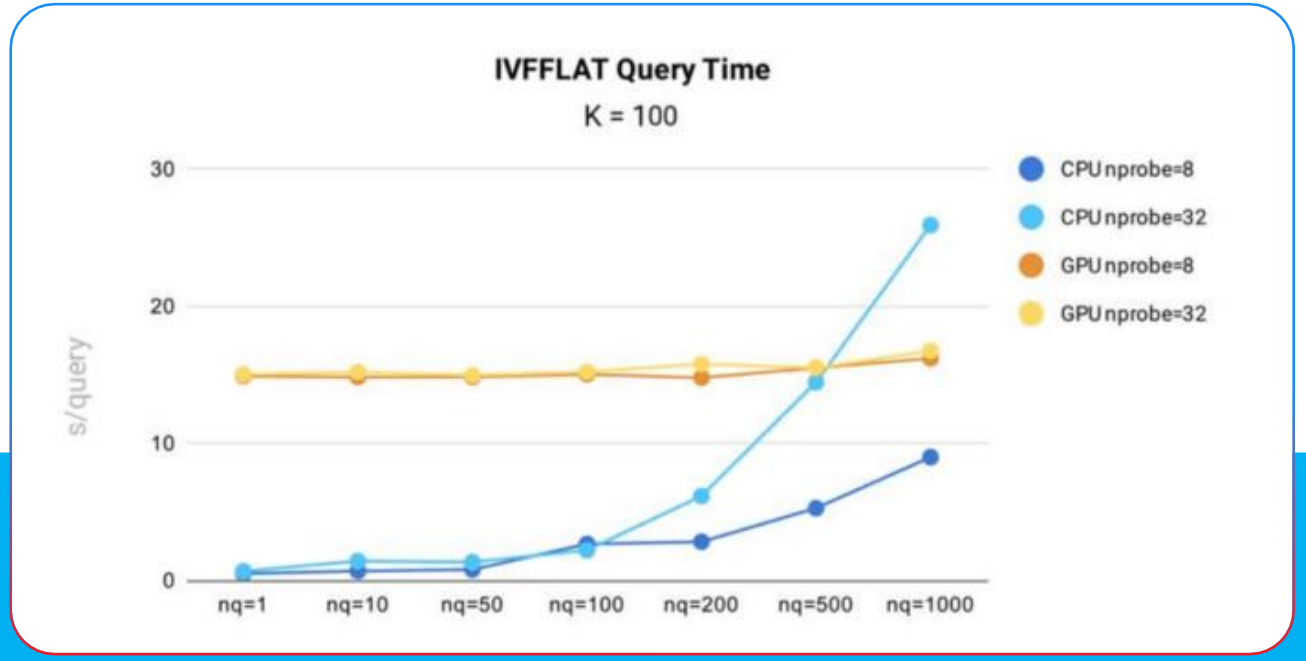


Milvus vector database

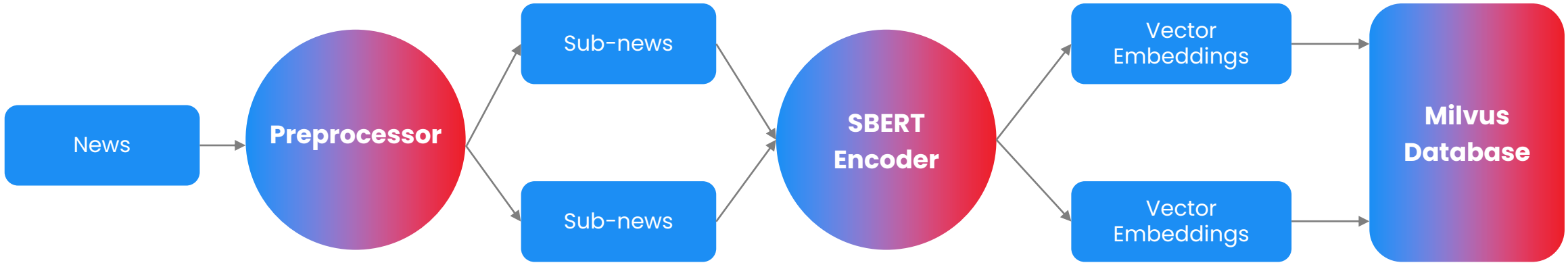
- Milvus uses the approximate nearest neighbor (ANN) search.
- “IVF_FLAT” index divides vector data into a number of cluster units (nlist) and compares distances between the target input vector and the center of each cluster.
- Depending on the number of clusters the system is set to query (nprobe), similarity search results are returned based on comparisons between the target input and the vectors in the most similar cluster(s) only — drastically reducing query time.

Milvus vector database

Query time test results for IVF_FLAT index in Milvus in Zilliz documentation on a dataset of 1 billion 128-dimensional vectors.



Evaluation

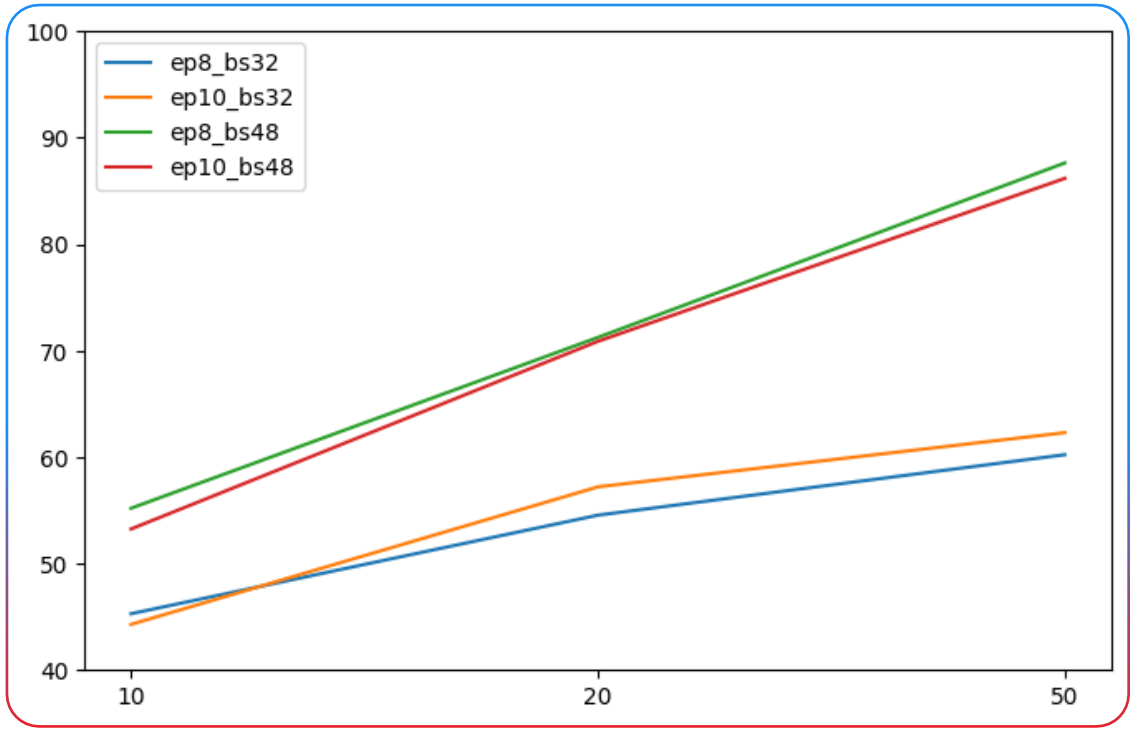


The pipeline for inserting knowledge base into Milvus Database

Evaluation

- At search phase, the model first vectorizes the input query.
- Then Milvus helps search that vector embedding in the knowledge base, returning k top-ranked results.
- This k value is often set to 100 for our use case if there are no advanced interference option.
- The naive evaluation process aim to maximize the precision with validation data set aside in the data preparation step.
- A search is evaluated as a success if the model can retrieve the ground truth label in the top K candidates.

Result



Retrieval rates at different epochs and batch sizes when retrieving from multiple top-k settings. ~15 inferences per second (query VNCoreNLP, encode with the model, search Milvus database)

Evaluation result

01

- We simply alter the parameters that affect model “fitness” on the dataset, like the number of training epochs, learning rates, and the optimizer; we did find improvements in loss values.
- At this time of research, we naively assumed that a better “fitness” means a better model, and we came up with the number of epochs to represent the model “fitness” on the dataset.

02

- We consider the causal reasoning factors. Increasing batch size can help the model’s reasoning ability because there will be more negative examples to compare.
- We found out that the effectiveness of increasing batch size exponentially reduced and accepted the value of 32 or 48.
- When we have similar sentences in the training data, it is impossible, not recommended, inconsequential to make the model learn to positive one over others.

FPT Education

FPT UNIVERSITY

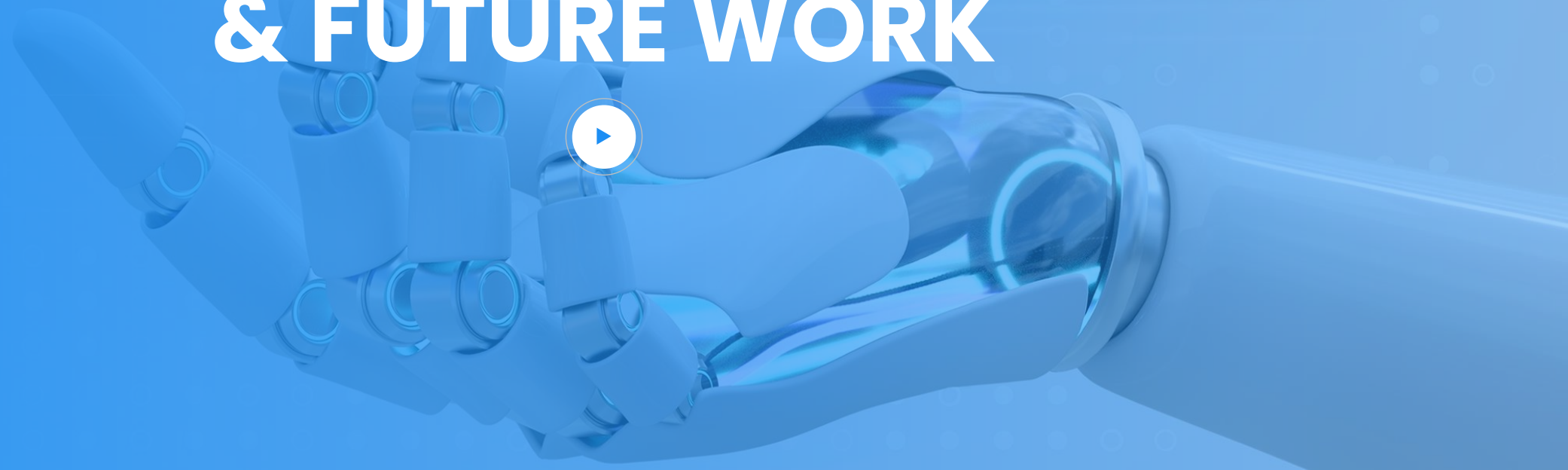
00

01

02

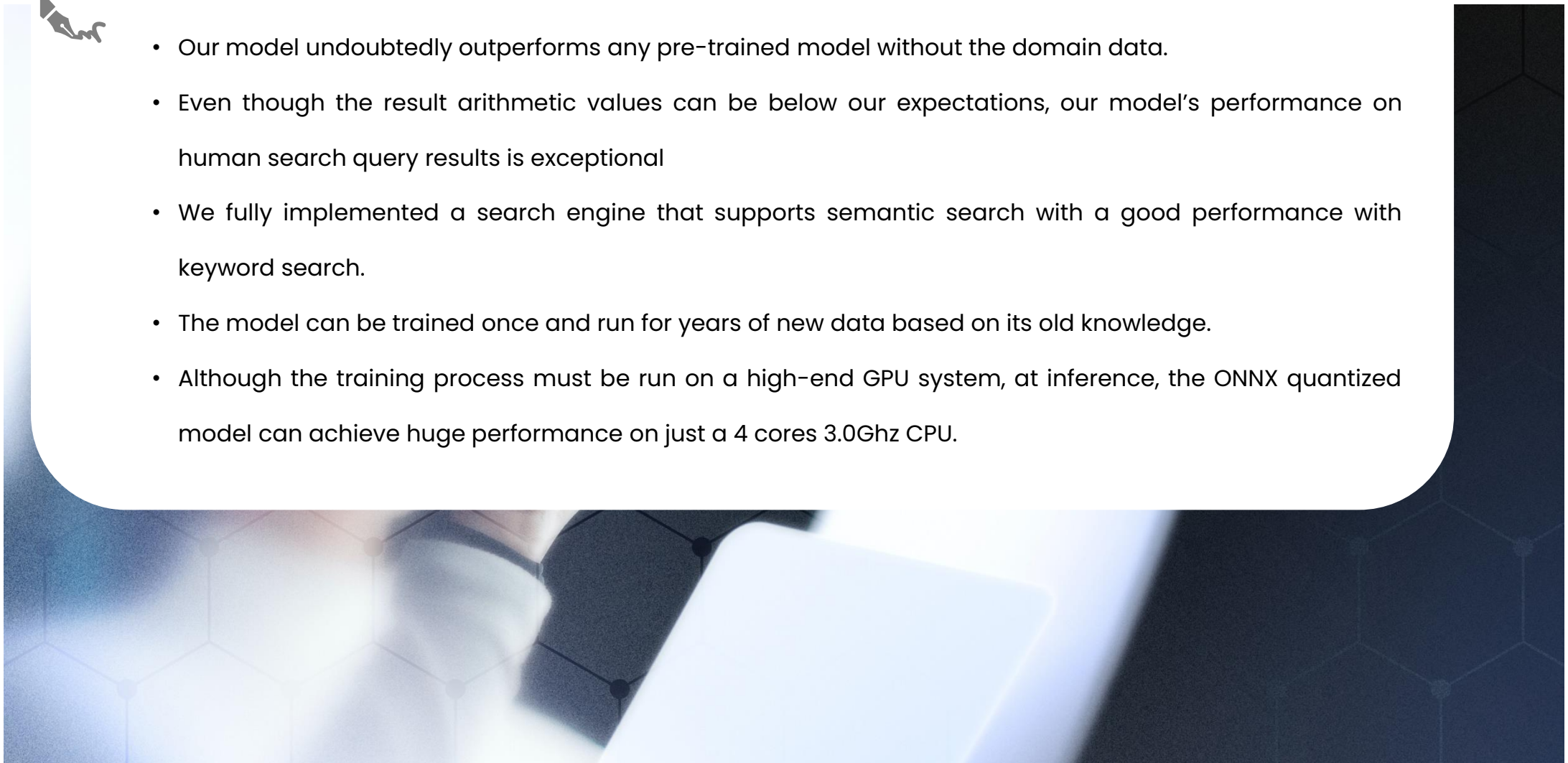
03

03 CONCLUSION & FUTURE WORK



Conclusion

- Our model undoubtedly outperforms any pre-trained model without the domain data.
- Even though the result arithmetic values can be below our expectations, our model's performance on human search query results is exceptional
- We fully implemented a search engine that supports semantic search with a good performance with keyword search.
- The model can be trained once and run for years of new data based on its old knowledge.
- Although the training process must be run on a high-end GPU system, at inference, the ONNX quantized model can achieve huge performance on just a 4 cores 3.0Ghz CPU.



THANK YOU!

Group

AIP490_G19