**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**KHỞI NGHIỆP KINH DOANH**

**BUSSINESS PLAN**

**ĐỀ TÀI: Auto-completion using Trie and Hybrid Approach of Sequential LSTM and N-gram**

| Sinh viên thực hiện: | | |
| --- | --- | --- |
| STT | Họ tên | MSSV |
| 1 | Phan Quốc Vỹ | 21522814 |
| 2 | Dương Tấn Hoàng | 21520866 |
| 3 | Nguyễn Duy Hưng | 21520894 |

GVHD: **ThS Nguyễn Trọng Chỉnh**

Lớp: **CS221.O11**

**TP. HỒ CHÍ MINH, tháng 01 năm 2024**

**LỜI CẢM ƠN**

Để hoàn thành đồ án môn học này, nhóm em xin chân thành cảm ơn đến:

Giảng viên lý thuyết - Thầy Nguyễn Trọng Chỉnh đã giảng dạy tận tình, chi tiết để chúng em có đủ kiến thức và vận dụng chúng vào đồ án này.

Giảng viên thực hành - Thầy Đặng Văn Thìn và Thầy Nguyễn Đức Vũ đã hướng dẫn, góp ý giúp chúng em hoàn thành đồ án này.

Do chưa có nhiều kinh nghiệm và hạn chế về kiến thức, lý luận của bản thân, kính mong nhận được những nhận xét, ý kiến đóng góp của các Thầy để đồ án môn học của chúng em được hoàn thiện hơn.

Nhóm chúng em xin trân trọng cảm ơn!

**MỤC LỤC**

[**1. GIỚI THIỆU 3**](#_pxyvmyefg02l)

[**2. NGỮ LIỆU 4**](#_krfjenlll6hb)

[2.1. Nguồn thu thập ngữ liệu 4](#_uvlecbxwcojd)

[2.2. Thống kê ngữ liệu 4](#_low3r9bd63fp)

[2.3. Phân tích một vài mẫu dữ liệu 6](#_vqed4huhsw8t)

[2.4. Nhận xét về ngữ liệu 20](#_a28aftyh7t3n)

[**3. TIỀN XỬ LÝ DỮ LIỆU 21**](#_do7ws1zby4yd)

[3.1. Tách bộ ngữ liệu 21](#_axfnldk8dt7p)

[3.2. Mã hóa các câu 21](#_e68j2a94isfy)

[3.3. Xử lý OOV 21](#_dt77czn2t1la)

[3.4. Phân chia bộ ngữ liệu 21](#_yiwg9fgk85pq)

[**4. PHƯƠNG PHÁP VÀ MÔ HÌNH 21**](#_lp2366p2gwt)

[4.1. Kiến trúc mô hình 21](#_j4yjdfacrikn)

[4.2. Hoàn thành từ 22](#_aooozuxic5p4)

[4.2.1. Cấu trúc dữ liệu Trie 22](#_kbh498tuhkrw)

[4.3. Dự đoán câu 23](#_mr1zqzbdnko)

[4.3.1. Mô hình Seq2Seq LSTM 23](#_g8fvpe9h5odb)

[4.4. Xếp hạng các câu dự đoán 24](#_jf26wvuazr77)

[4.4.1. Mô hình N-gram 24](#_4z7sz5xyurl6)

[**5. ĐÁNH GIÁ 25**](#_6jlba1q62zm5)

[5.1. Thang đo BLEU 25](#_60xxwjq4xxtd)

[5.2. Thang đo ROUGE 26](#_usj97dkcy40v)

[5.3. Thang đo BERT 27](#_y5nf29sgeirl)

[**TÀI LIỆU THAM KHẢO 28**](#_v3rz0xcqj0r9)

# GIỚI THIỆU

Trong các hệ thống liên quan đến truy vấn hoặc giao diện dựa trên đề xuất, tự động hoàn thành và dự đoán chuỗi có thể là một tính năng cốt lõi để tích hợp hệ thống với tương tác người dùng. Điều này cải thiện tương tác người dùng bằng cách đoán các từ tiếp theo mà người dùng muốn gõ.

Dự đoán chuỗi có nghĩa là đoán từ tiếp theo, thậm chí cả câu hoàn chỉnh từ chuỗi từ trước đó. Cụ thể hơn, dự đoán chuỗi là một quá trình xử lý dự đoán giá trị hoặc giá trị tiếp theo trong chuỗi dựa trên các giá trị trước đó. Chuỗi này có thể là các chữ cái hoặc từ trong một câu. Tự động hoàn thành giúp người dùng hoàn thành từ, giảm thiểu lỗi chính tả. Tự động hoàn thành và dự đoán chuỗi được cho là một tính năng thiết kế giao diện người dùng tốt hơn cho phần mềm. Dự đoán chuỗi thuộc về vấn đề học máy giám sát. Có một số phương pháp để cải thiện giải pháp. Các nghiên cứu trước đây về lĩnh vực này đã đưa ra ánh sáng, tuy nhiên, hệ thống dựa trên dự đoán trong ngôn ngữ Anh chưa đạt được mức độ hài lòng đó. Hiện nay, chúng ta có hàng ngàn phần mềm và ứng dụng dựa trên đề xuất hoặc Q/A sử dụng ngôn ngữ Anh với hiệu suất tự động hoàn thành hoặc dự đoán ngữ cảnh người dùng thấp hơn. Với sự tập trung ngày càng tăng của xử lý ngôn ngữ tự nhiên trong ngôn ngữ Anh và việc triển khai của nó, lĩnh vực này thu hút sự chú ý. Một số công trình nghiên cứu đề xuất việc triển khai mô hình N-gram với sự loại bỏ các nội suy và mô hình backoff để dự đoán từ tiếng Anh. Một số nghiên cứu cũng tập trung vào việc triển khai Mạng Nơ-ron cho dự đoán từ tiếng Anh. Mô hình N-gram có thể phù hợp cho dự đoán từ. Tuy nhiên, để hiệu quả hơn và phù hợp với các loại dữ liệu khác nhau, các phương pháp Mạng Nơ-ron đã được triển khai trong công trình này. Toàn bộ khung hệ thống có thể được chia thành hai phần:

* **Word Completion**: Sử dụng cấu trúc dữ liệu Trie để lưu trữ từ điển được tạo từ bộ dữ liệu.
* **Sequence prediction**: Kết hợp 2 mô hình là Sequential LSTM và N-gram. Trong đó, hệ thống dự đoán sẽ tập trung vào sử dụng mô hình Sequential LSTM cụ thể là mô hình Seq2Seq để tạo ra một danh sách gồm 5 câu dự đoán. Sau đó sẽ sử dụng mô hình N-gram để sắp xếp thứ tự của danh sách này dựa trên xác suất của mỗi câu.

Thông qua việc xem xét một số công trình trước đây về dự đoán chuỗi tiếng Anh, chúng em nhận thấy rằng hầu hết các công trình này tập trung vào việc sử dụng mô hình N-gram để dự đoán từ tiếp theo. Lý do sử dụng mô-đun LSTM với N-gram là vì LSTM là một mô hình mạnh mẽ và linh hoạt . Ngoài ra, đối với các trường hợp chưa từng thấy, nó phù hợp một cách trơn tru và mang lại kết quả đáng hài lòng. Có một số tác động của khung hệ thống đề xuất có thể áp dụng cho bất kỳ tính năng dựa trên tự động hoàn thành nào trong tiếng Anh, bàn phím, chatbot, hệ thống đề xuất, hệ thống dựa trên Q/A, v.v.

# NGỮ LIỆU

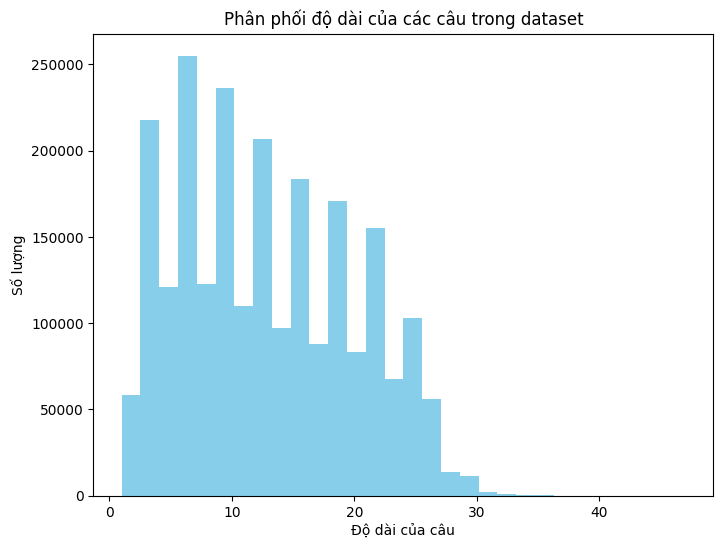
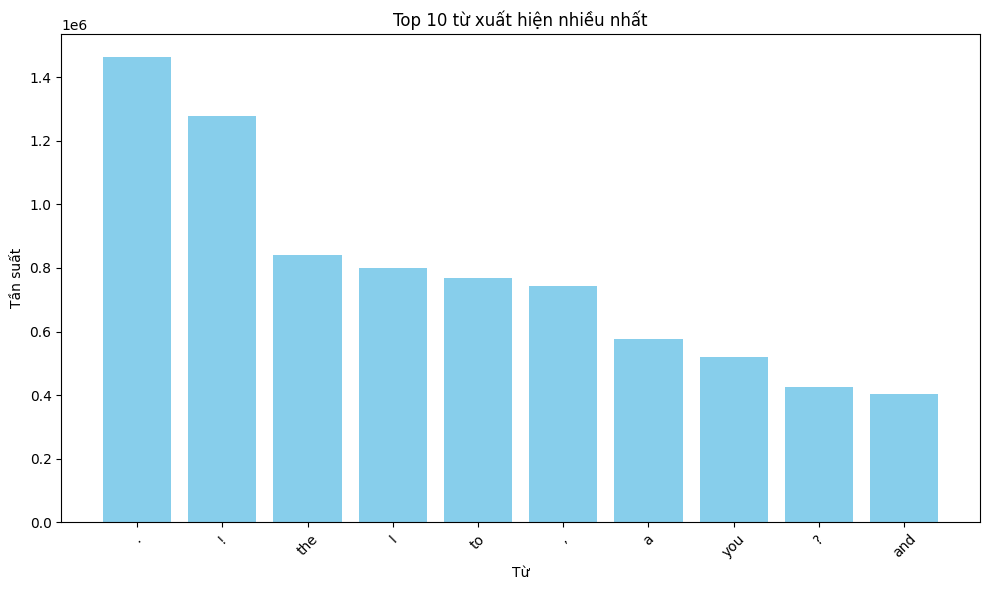
## Nguồn thu thập ngữ liệu

* Sử dụng bộ dữ liệu có sẵn cho bài toán Autocomplete từ Kaggle:

<https://www.kaggle.com/code/sauravmaheshkar/auto-completion-using-n-gram-models/input>

* Bộ dữ liệu chứa các dòng văn bản thu thập từ Twitter trong tiếng Anh (US) Dữ liệu này có thể bao gồm các bài đăng, tweet, thông điệp ngắn hoặc các phần mô tả ngắn về các chủ đề khác nhau. Các dòng văn bản trong tập dữ liệu này có thể bao gồm ngôn ngữ tự nhiên, các từ viết tắt, biểu cảm, emoji và có thể chứa thông tin về các sự kiện, ý kiến cá nhân, hoạt động hàng ngày, tin tức, sự kiện văn hóa, thời sự và nhiều chủ đề khác. Dữ liệu từ Twitter có thể mang tính cách mạng và đa dạng, phản ánh ý kiến và suy nghĩ của người dùng Twitter từ nhiều quan điểm và góc độ khác nhau.

## Thống kê ngữ liệu

* Bộ dữ liệu bao gồm **164.456.394** kí tự nhưng vì giới hạn về phần cứng, nhóm chỉ có thể sử dụng **500.000** ký tự đầu tiên tương ứng với **7214** đoạn văn bản cho thực nghiệm
* Thống kê độ dài các đoạn văn bản trong dataset:  
  
* Thống kê 10 ký tự xuất hiện nhiều nhất trong dataset:  
  
* WordCloud của dataset:  
  

## Phân tích một vài mẫu dữ liệu

* Phân tích 70 câu dữ liệu:

Câu 1:

"congrats to all who made last night a sellout .. come"

Câu trên là một câu chúc mừng và mời gọi những người góp phần bán hết vé đến. "Congrats" là viết tắt của "congratulations" (chúc mừng). Câu trên chỉ là một câu bình thường. Chỉ có một chỗ đặc biệt ở dấu “..”, đây có thể là lỗi sai hoặc mang ý nghĩa ngại ngùng, ngắt đoạn trong câu nói.

Câu 2:

"oh - thank you so much : -d"

Câu trên là một câu cảm ơn. Điều đặc biệt nằm ở icon ở cuối câu biểu hiện gương mặt rạng rỡ, hài lòng.

Câu 3:

"please pardon my timeline for the next hour . i am joining monthly chat . join us to learn more ."

Ở câu trên tác giả muốn xin lỗi vì thời gian sắp tới tác giả sẽ bận tham gia cuộc trò chuyện hàng tháng và mời gọi mọi người cùng tham gia. Câu trên là một câu đúng ngữ pháp và không có điểm đặc biệt.

Câu 4:

"fact : typewriter is the longest english word that can be made using only the letters on the top row of a computer keyboard ."

Câu trên là một câu nói lên một sự thật hiển nhiên rằng từ “typewriter” là từ dài nhất có thể được tạo ra bởi hàng trên cùng của bàn phím máy tính. Câu trên không có gì đặc biệt và đã viết đúng ngữ pháp.

Câu 5:

"seriously ? i 'll have to check it out !"

Câu trên mang thái độ ngạc nhiên hoặc hoài nghi và cho biết rằng họ sẽ kiểm tra điều đó để xác nhận hoặc tìm hiểu thêm thông tin. Dấu chấm than (!) sau "seriously" thể hiện sự ngạc nhiên hoặc thắc mắc. Với một câu hỏi rút gọn và xuất hiện một từ viết tắt “will” thành “‘ll”.

Câu 6:

"sid is back"

Câu trên là một câu thông báo ngắn gọn, đúng ngữ pháp và không có gì đặc biệt.

Câu 7:

"# foodhubs are becoming key aggregation point for traditional distributors ( so cool for producers ! )"

Câu trên là một câu thông báo, đính kèm với đó là một bình luận được viết ở trong ngoặc đơn. Điểm đặc biệt của câu này là có phần giải thích được viết bên trong dấu ngoặc đơn.

Câu 8:

"couples jeggings ! brilliant !"

Câu trên có thể hiểu là người nói đang bày tỏ sự thích thú hoặc ngạc nhiên với ý tưởng hoặc sản phẩm "couples jeggings" (jeggings dành cho cặp đôi). Từ "brilliant" (tuyệt vời) được sử dụng để diễn đạt sự ấn tượng hoặc hài lòng đối với ý tưởng đó.

Câu 9:

"sometimes you got ta do stuff you really do n't wan na do ."

Câu trên diễn tả ý nghĩa rằng đôi khi chúng ta phải thực hiện những nhiệm vụ mà thực sự không muốn làm.

Câu 10:

"really craving some taco bell ! idk even know why-\_-"

Câu trên diễn tả sự khao khát ăn Taco Bell và người nói không hiểu tại sao họ lại cảm thấy như vậy. Ký hiệu "-\_-" thường được sử dụng để biểu lộ sự khó chịu hoặc cảm xúc tiêu cực.

Câu 11:

"watch # greatdaysa today and see what makes a 18 year old ride his bike 30,000 miles across 2 continents to break 5 world records !"

Câu trên là một lời giới thiệu và mời người nghe xem chương trình #greatdaysa hôm nay. Nó nhấn mạnh về một câu chuyện đặc biệt về một người trẻ tuổi đi xe đạp hàng ngàn dặm để phá vỡ nhiều kỷ lục thế giới.

Câu 12:

"im tweetin on the downsides of these alpine climbs ! ! when done here .. im gon na develop a # westernskillet flavored"

Câu trên khá phức tạp cũng như xuất hiện lỗi sai chính tả ở “từ tweetin”. Sử dụng 2 dấu “!”, sử dụng từ nhấn mạnh, viết tắt ở “im”, sử dụng “gonna”. Tóm lại câu trên là một câu rất khó để huấn luyện cũng như đoán ra; có thể xem câu trên như một test nhiễu.

Câu 13:

"ncl has all sailings from ny with reduces deposit free onboard credit sale ends monday details jackie 877-416-9722 ext 4413"

Câu trên là một thông báo quảng cáo về các chuyến đi du thuyền từ công ty NCL. Nó cho biết về chính sách giảm tiền đặt cọc và tín dụng sử dụng trên tàu miễn phí. Câu cũng cung cấp thông tin liên hệ để biết thêm chi tiết.

Câu 14:

"oh indeed , this will be great !"

Câu trên cho thấy người nói đồng ý và rất hào hứng với thông tin được cung cấp trước đó. Câu trên ngắn gọn, đúng ngữ pháp.

Câu 15:

"same here might order it tonight or tomorrow cause will have to get it on brads card hehe x"

Câu trên diễn đạt sự đồng ý và chia sẻ ý kiến. Người nói định ý định đặt hàng vào tối nay hoặc ngày mai và sẽ sử dụng thẻ của Brad để thanh toán. Biểu tượng "hehe x" thể hiện một sự vui vẻ hoặc cười.

Câu 16:

"lo mejor ! it was cool experiencing work through the eyes of an excited teen"

Câu trên diễn đạt sự hài lòng và khen ngợi về việc trải nghiệm công việc từ góc nhìn của một cô hoặc cậu thiếu niên hào hứng. Từ "lo mejor" trong tiếng Tây Ban Nha thể hiện sự hài lòng và ấn tượng tích cực.

Câu 17:

"fuck working out without headphones . i should 've just gone home"

Câu trên diễn đạt sự tức giận và hối tiếc vì tập thể dục mà không có tai nghe. Từ ngữ tục tĩu "fuck" được sử dụng để diễn đạt sự tức giận, thất vọng hoặc không hài lòng.

Câu 18:

"spurs did n't play good on d , parker had an off night shooting ... just saying"

Câu trên diễn đạt ý kiến rằng Spurs đã không chơi tốt trong phòng ngự (d viết tắt cho defence) và Parker có một đêm không chính xác trong việc ghi điểm. Người nói chỉ đơn thuần đưa ra thông tin và ý kiến của mình (just saying). Câu trên tương đối khó để nhận ra do có từ viết tắt và ngữ nghĩa phải phụ thuộc vào vế sau.

Câu 19:

"only a few nights left of marc trio ! ! ! !"

Câu trên diễn đạt sự hứng thú hoặc sự tiếc nuối về việc chỉ còn vài đêm nữa của "Marc trio". Cụm từ "Marc trio" có thể ám chỉ một nhóm người có tên là Marc hoặc một sự kết hợp gì đó liên quan đến Marc. Nhiều dấu “!” thể hiện cảm xúc hưng phấn.

Câu 20:

"surpassed our # grizzgala fundraising auction goal ! building the ouaa scholarship endowment , thx to sponsors & the great auction donations !"

Câu trên diễn đạt sự vui mừng và cảm ơn vì đã vượt qua mục tiêu gây quỹ từ cuộc đấu giá #grizzgala. Người nói cũng đề cập đến việc xây dựng quỹ học bổng ouaa và cảm ơn các nhà tài trợ và các đóng góp đồ đấu giá tuyệt vời. Câu trên xuất hiện từ viết tắt “thx” và ký tự đặc biệt “&”.

Câu 21:

"i love ignoring people , but i hate being ignored ."

Câu trên diễn đạt một mâu thuẫn trong suy nghĩ hoặc cảm xúc của người nói. Họ thích không để ý đến người khác, nhưng đồng thời cảm thấy không vui khi bị bỏ qua hoặc không được chú ý.

Câu 22:

"holy black mamba !"

Câu trên là một câu thốt lên hoặc biểu đạt sự ngạc nhiên, kinh ngạc hoặc thán phục. Câu trên xuất hiện hình ảnh ẩn dụ “black mamba” thể hiện sự không thể tin được.

Câu 23:

"“ : i shall be listening to beastie boys all day # ripmca ” < -- another reason rory kicks ass ."

Câu trên diễn đạt rằng người nói sẽ dành cả ngày để nghe nhạc của nhóm Beastie Boys và sử dụng hashtag #ripmca để tưởng nhớ và tôn vinh Adam Yauch.

Câu 24:

"oh , by all means ! yes , ma'am !"

Câu trên diễn đạt sự đồng ý hoặc sẵn lòng tuân theo yêu cầu của người khác. Từ "oh, by all means!" diễn đạt sự đồng ý và khích lệ, trong khi "yes, ma'am!" diễn đạt sự tôn trọng và sẵn lòng đáp ứng yêu cầu của người phụ nữ.

Câu 25:

"i will happily provide cookies in exchange for answers . very sad to hear about # awake you are an exceptional writer ."

Câu trên diễn đạt sự sẵn lòng cung cấp bánh quy để đổi lấy câu trả lời. Ngoài ra, người nói cũng diễn đạt sự buồn bã khi nghe tin tức liên quan đến sự việc được gắn với hashtag "#awake". Người nói cũng ca ngợi người viết bài là một nhà văn hoặc biên kịch xuất sắc.

Câu 26:

"by the color they made his skin look , its"

Câu trên biểu đạt sự giải thích màu làn da của người đàn ông. Câu trên thực hiện câu rút gọn và khá khó hiểu hay dịch.

Câu 27:

"you 're welcome = ) < 3"

Câu trên diễn đạt sự đáp lại lời cảm ơn và biểu đạt sự vui mừng và tình cảm. Người nói cho biết họ đã nhận được lời cảm ơn và gửi đi sự chân thành và thiện chí. Câu trên xuất hiện icon “=)” và “<3”.

Câu 28:

"« i want a large cherry berry chiller . »"

Câu trên diễn đạt mong muốn của người nói muốn có một ly Cherry Berry Chiller lớn. Câu trên đúng ngữ pháp.

Câu 29:

"gon na be a long day"

Câu trên diễn đạt rằng ngày hôm đó sẽ kéo dài và có thể mệt mỏi hoặc căng thẳng. Câu trên sử dụng từ “gonna” không thực sự đúng ngữ pháp nhưng khá thông dụng.

Câu 30:

"nothing is certain outside of yourself ."

Câu trên diễn đạt ý rằng không có gì chắc chắn ngoài chính bản thân mình. Câu trên là một câu châm ngôn, đúng ngữ pháp.

Câu 31

"# 3wordsforyou trust in god "

Câu trên khuyến khích việc đặt niềm tin và sự tin cậy vào Chúa. Hashtag #3wordsforyou cho thấy yêu cầu sử dụng ba từ để truyền đạt một thông điệp.

Câu 32:

"hey giants fans . how ya feeling tonight . boy , do you look dumb in that stupid panda hat ."

Câu trên chứa sự chào hỏi đối với người hâm mộ đội Giants và hỏi cảm nhận của họ vào buổi tối đó. Tuy nhiên, câu kết thúc bằng một phê phán tiêu cực với việc đánh giá rằng người mặc chiếc mũ panda trông ngớ ngẩn.

Câu 33:

"what am i doing this morning ? taking photos at a fencing tournament !"

Câu trên diễn đạt rằng vào buổi sáng đó, người nói sẽ tham gia vào việc chụp ảnh tại một giải đấu đấu kiếm.

Câu 34:

"hey whats up ?"

Câu trên là một cách chào hỏi thông thường và thân mật. Người nói đang tìm hiểu về tình hình hiện tại của người nghe, muốn biết những gì đang xảy ra hoặc những gì đang làm. Đây là một câu hỏi mở, mời người khác chia sẻ về tình hình của họ và tạo cơ hội cho một cuộc trò chuyện tiếp theo.

Câu 35:

"you 're welcome ! enjoy cc rt : sweet potato # popchips are amazeballs ! ! thanks for the vday treats"

Câu trên diễn đạt sự đáp lại lời cảm ơn và chia sẻ sự thích thú với sản phẩm Popchips hương khoai lang. Người nói cảm ơn vì nhận được những món quà ngọt trong dịp Valentine. Câu cũng có thể đề cập đến Sweet Potato như một người hoặc thương hiệu liên quan đến Popchips.

Câu 36:

"i like how my room stays at comfortable 17 million degrees : d"

Câu trên diễn đạt một sự phóng đại đùa vui về nhiệt độ trong phòng ngủ của người nói. Bằng cách nói "17 triệu độ", người nói nhấn mạnh rằng phòng của họ rất ấm hoặc nóng. Sử dụng biểu tượng cảm xúc ":D" cho thấy một tinh thần đùa vui và hài hước.

Câu 37:

"karen and drita do n't have to be friends . they should squash it so they could at least be in the same room"

Câu trên diễn đạt ý kiến rằng Karen và Drita không cần phải là bạn nhưng cần giải quyết vấn đề giữa họ để có thể chung sống trong cùng một không gian. Điều này ngụ ý rằng có một xung đột hoặc mâu thuẫn giữa hai người. Người nói mong muốn họ giải quyết và hòa giải để tạo ra một môi trường thoải mái hơn, nơi họ có thể ở chung mà không có căng thẳng hoặc thù địch.

Câu 38:

"more skating ! come by the check out a movie , eat a great dinner and top it off with great times at the ice rink ."

Câu trên diễn đạt sự hào hứng với việc trượt băng và mời người khác tham gia. Người nói đề nghị xem một bộ phim, có một bữa tối ngon và kết thúc ngày vui với những khoảnh khắc đáng nhớ tại sân trượt băng. Sử dụng dấu chấm than (!) để truyền tải sự phấn khích và sự mong đợi.

Câu 39:

"i can never tell with you .."

Câu trên diễn tả sự kiên quyết trong việc giữ một bí mật gì đó. Câu trên đúng ngữ pháp và có 2 dấu “.” cuối câu có lẽ do người viết nhầm.

Câu 40:

"how do you keep from getting sick with all of this weather change ?"

Câu trên diễn đạt sự tìm kiếm lời khuyên để duy trì sức khỏe trong điều kiện thời tiết thay đổi liên tục.

Câu 41:

"o .. intrigued by ! we 're in negotiations for building our new website ..."

Câu trên diễn đạt sự hào hứng và quan tâm đối với việc đàm phán xây dựng một trang web mới.

Câu 42:

"the minnesota twins have agreed to terms with pitcher francisco liriano on a one-year , $ 4.3 million contract , avoiding arbitration ."

Câu trên cung cấp thông tin về việc đội bóng Minnesota Twins đã đạt được thỏa thuận với nghệ sĩ ném bóng Francisco Liriano về một hợp đồng một năm trị giá 4,3 triệu đô la.

Câu 43:

"proud to have voted twice for the sag-aftra merger . have you voted yet ?"

Câu trên diễn đạt sự tự hào của người viết vì đã tham gia vào việc bỏ phiếu hai lần để ủng hộ việc sáp nhập giữa hai tổ chức SAG và AFTRA. Đồng thời, người viết cũng hỏi người đọc xem họ đã tham gia bỏ phiếu chưa.

Câu 44:

"another apology from david letterman - there 's talk he may be punished . do you think he should be disciplined by the network ? let us know ..."

Câu trên đề cập đến việc David Letterman đã xin lỗi một lần nữa và có những đồn đoán về việc anh ta có thể bị trừng phạt. Người nói sau đó yêu cầu ý kiến về việc liệu anh ta có nên bị trừng phạt bởi đài phát thanh, truyền hình hoặc tổ chức mà anh ta đang làm việc.

Câu 45:

"sorry to know you missed the event . hope to see you at the restaurant soon : )"

Câu trên diễn đạt sự lấy làm tiếc vì người nghe đã không tham gia sự kiện cụ thể nào đó. Người nói thể hiện mong muốn gặp lại người nghe tại nhà hàng trong tương lai gần. Câu này mang tính chất lịch sự và thể hiện sự hy vọng về một cuộc gặp gỡ trong tương lai.

Câu 46:

"you 're one of them . thanks john ."

Câu trên diễn tả sự khẳng định cũng như sự biết ơn của người dùng với nhân vật “john”. Câu trên đúng ngữ pháp, xuất hiện từ rút gọn “`re”.

Câu 47:

"you 'd think trained librarians with master 's degrees would n't have such a hard time figuring out how to unsubscribe from a listserv . jeez ."

Câu trên diễn đạt sự ngạc nhiên và thất vọng của người nói về việc các thủ thư đã qua đào tạo và có bằng thạc sĩ lại gặp khó khăn trong việc hủy đăng ký khỏi một listserv. "Jeez": Từ này thường được sử dụng để diễn đạt sự ngạc nhiên, thất vọng hoặc tức giận.

Câu 48:

"you bet . now how do you land this crazy thing ?"

Câu trên có ý nghĩa là người nói đồng ý với điều gì đó hoặc đề xuất trước đó. Sau đó, người nói hỏi về cách hạ cánh một "crazy thing" mà không đưa ra thông tin cụ thể về vấn đề đang được thảo luận.

Câu 49:

"vinny and his spelling mistakes ! lol 😁 it 's cute 😊"

Câu trên thể hiện sự vui nhộn và thích thú của người nói về những lỗi chính tả của Vinny. Họ cho rằng những lỗi đó là dễ thương và đáng yêu bằng việc sử dụng "lol", "😁" và "😊" .

Câu 50:

"wordcamp ? there is one in milwaukee too ."

Câu trên cho thấy người nói đã nghe về WordCamp và diễn đạt sự ngạc nhiên hoặc tò mò về sự tồn tại của nó. Sau đó, người nói chia sẻ thông tin rằng có một WordCamp diễn ra ở Milwaukee.

Câu 51:

"1/4/12 : today 's pink slip topic is based on a favorite quote from the book `` the doors we open and close each day decide the lives we live ''"

Câu trên là một câu thông báo về chủ đề của ngày hôm 1/4/12. Câu trên viết đúng ngữ pháp và không có gì đặc biệt.

Câu 52:

"i would love to know who thot that opening a recorded call w/ `` do not hang up '' was a good way to get someone to do anything but that ."

Câu trên diễn đạt sự bất ngờ và không hiểu của người nói về cách mà một cuộc gọi đã được ghi âm mở đầu bằng cách cảnh báo "đừng treo máy". Người nói cho rằng việc sử dụng cụm từ này không phải là một cách hiệu quả để thuyết phục ai đó làm điều gì đó ngoài việc treo máy.

Câu 53:

"sonny rollins honored on cbs : kennedy centers honors ! jazz on network television . that 's real `` reality '' television !"

Câu trên thể hiện sự phấn khích và thích thú của người nói về việc Sonny Rollins được tôn vinh trong chương trình Kennedy Center Honors trên mạng truyền hình CBS. Người nói cũng thể hiện sự vui mừng với việc có một chương trình về jazz được phát sóng trên mạng truyền hình, coi đó là một ví dụ thực sự về "thực tế" trên truyền hình.

Câu 54:

"i think it 's interesting . but i might be biased : )"

Câu trên cho thấy người nói thấy thú vị về chủ đề nào đó. Tuy nhiên, họ cũng nhận thức rằng quan điểm của mình có thể bị thiên vị.

Câu 55:

"the thing i wish i could say to you but knowing i wont < < <"

Câu trên có vẻ như đang diễn tả một cảm xúc hoặc suy nghĩ mà người nói muốn chia sẻ với ai đó nhưng biết rằng họ sẽ không làm điều đó. Cụm từ “< < <” có thể được hiểu là một dấu hiệu của sự ngần ngại, hoặc một cách để chỉ ra rằng có điều gì đó không được nói ra. Tuy nhiên, nó cũng có thể phụ thuộc vào ngữ cảnh cụ thể mà câu này được sử dụng. Câu này được viết đúng ngữ pháp và không có gì đặc biệt.

Câu 56:

"thanks for giving us a try ! hope you enjoy your food"

Câu trên là một câu cảm ơn khách hàng đã thử dịch vụ hoặc sản phẩm của họ, đồng thời cũng hy vọng khách hàng sẽ thích món ăn. Câu này được viết đúng ngữ pháp và không có gì đặc biệt. Dấu chấm than (!) sau “thanks for giving us a try” và “hope you enjoy your food” thể hiện sự phấn khích và lòng nhiệt tình.

Câu 57:

"do n't know if i can trust you ."

Câu trên diễn đạt một cảm xúc hoặc suy nghĩ mà người nói không chắc chắn về việc có thể tin tưởng vào người khác hay không. Câu này được viết đúng ngữ pháp và không có gì đặc biệt.

Câu 58:

"yup , we 're just off the 5 and on katella . thanks for checking us out , we 'll be sure not to disappoint ."

Câu trên có vẻ như đang chỉ ra vị trí của một địa điểm, có thể là một cửa hàng hoặc nhà hàng, nằm gần đường số 5 và trên đường Katella. Người nói cảm ơn vì đã kiểm tra thông tin về họ và hứa sẽ không làm thất vọng. Câu này được viết đúng ngữ pháp và không có gì đặc biệt. Câu này cũng sử dụng từ viết tắt “will” thành “‘ll”.

Câu 59:

"thanks for the retweet , maria !"

Câu trên là một lời cảm ơn dành cho Maria vì đã chia sẻ lại (retweet) một bài đăng nào đó. Câu này được viết đúng ngữ pháp và không có gì đặc biệt. Dấu chấm than (!) sau “thanks for the retweet, Maria” thể hiện sự phấn khích và lòng biết ơn.

Câu 60:

"go see it and see for yourself why you should n't go see it . - samuel goldwyn"

Câu trên là một trích dẫn của Samuel Goldwyn, một nhà sản xuất phim nổi tiếng. Câu này có vẻ như đang mô tả một sự mâu thuẫn hài hước: khuyến khích mọi người xem một thứ gì đó để hiểu tại sao họ không nên xem nó. Đây có thể được hiểu là một cách châm biếm hoặc chỉ trích một cách duyên dáng về một sản phẩm nào đó, có thể là một bộ phim. Câu này được viết đúng ngữ pháp và không có gì đặc biệt.

Câu 61:

"she 's gon na be loads of fun by wedding time . loads ."

Câu trên có vẻ như đang nói về một người phụ nữ sẽ rất vui vẻ vào thời điểm đám cưới. Từ “loads” được lặp lại để nhấn mạnh mức độ vui vẻ hoặc hứng thú. Câu này được viết đúng ngữ pháp và không có gì đặc biệt. Câu này cũng sử dụng từ viết tắt “going to” thành “gon na”.

Câu 62:

"hey im bored lol"

Câu này không có gì đặc biệt và không có lỗi ngữ pháp. lol viết tắt cho laughing out loud, nghĩa là mắc cười.

Câu 63:

"# teamyandy chrissy ah old ass bitch w/ zero class"

Câu trên là một câu lăng mạ được viết với thái độ tức giận. Ở câu trên xuất hiện w/ có lẽ mang ý nghĩa như từ with.

Câu 64:

"they be boppin on a handstand"

Câu trên có thể được hiểu là một người hoặc một nhóm người đang thực hiện một động tác boppin (một kiểu nhảy múa) trong khi đứng bằng tay. Đây có thể là một cách diễn đạt hình ảnh hoặc hành động một cách sáng tạo và hài hước. Tuy nhiên, câu này không tuân thủ hoàn toàn ngữ pháp tiếng Anh chuẩn.

Câu 65:

"our hosting provider is having some problems this morning , but we should be back up shortly . i 'm sorry for the inconvenience !"

Câu này được viết đúng ngữ pháp và mang tính chất chính thức, thích hợp khi giao tiếp với khách hàng hoặc người dùng. Dấu chấm than (!) sau “inconvenience” nhấn mạnh sự xin lỗi và lòng thành của người nói.

Câu 66:

"awful managing in the ninth , ron . calling for a sacrifice with the worst bunter on the team is not a good call . horrible ."

Câu trên là một phản hồi tiêu cực về quản lý của Ron trong hiệp chín của một trận đấu (có thể là bóng chày). Người nói không hài lòng với quyết định của Ron khi yêu cầu một cú hy sinh từ người chơi có kỹ năng bunt kém nhất trong đội. Câu này được viết đúng ngữ pháp và mang tính chất phê phán. Dấu chấm than sau “horrible” nhấn mạnh sự không hài lòng và thất vọng của người nói.

Câu 67:

"you are very welcome ! ! thank you as well !"

Câu trên là một câu trả lời phổ biến trong giao tiếp hàng ngày. Người nói đang bày tỏ lòng biết ơn và cảm kích, đồng thời cũng chào đón người khác. Câu này được viết đúng ngữ pháp và mang tính chất thân thiện, lịch sự. Dấu chấm than kép sau “welcome” và dấu chấm than sau “well” nhấn mạnh sự nhiệt tình và lòng biết ơn của người nói.

Câu 68:

"anyone who wants to make a difference and give of themselves should start training for the # peaceride ! it 's going to rock & help !"

Câu trên là một lời kêu gọi cho những ai muốn tạo ra sự khác biệt và dành thời gian, công sức của mình nên bắt đầu tập luyện cho sự kiện #peaceride. Người nói tin rằng sự kiện này sẽ tuyệt vời (rock) và có ích (help). Câu này được viết đúng ngữ pháp và mang tính chất khích lệ, động viên. Dấu chấm than sau “peaceride” và “help” nhấn mạnh sự hào hứng và tính quan trọng của sự kiện.

Câu 69:

"you should probably include some of our sweet tea vodka in that arnold palmer . we wo n't tell on ya !"

Câu trên là một lời đề nghị thêm một chút vodka trà ngọt vào đồ uống Arnold Palmer. Câu này được viết đúng ngữ pháp và mang tính chất giao tiếp không chính thức. Dấu chấm than sau “palmer” và “ya” nhấn mạnh sự hào hứng và tính cách thân thiện của người nói.

Câu 70:

"no it 's okay ! do n't feel bad !"

Câu trên là một câu trấn an, khuyên nhủ người khác không nên cảm thấy tệ hại. Câu này được viết đúng ngữ pháp và mang tính chất thân thiện, an ủi. Dấu chấm than sau “okay” và “bad” nhấn mạnh sự quan tâm và lòng tốt của người nói.

## Nhận xét về ngữ liệu

Ngữ liệu trên đa phần là các đoạn văn bản, các câu văn bản ngắn. Đoạn dài nhất có 39 từ, đoạn ngắn nhất có 2 từ. Xuất hiện đa dạng các kiểu cấu trúc câu, các ký tự đặc biệt, các lỗi chính tả.

Tóm lại bộ dữ liệu tương đối đa dạng và cấu trúc phức tạp. Tương đối khó để thực hiện huấn luyện.

# TIỀN XỬ LÝ DỮ LIỆU

## Tách bộ ngữ liệu

Đầu tiên là thực hiện lấy ra 500.000 ký tự, sau đó thực hiện cắt ra các từ cách nhau bằng dấu cách “ ”.

## Mã hóa các câu

Có thể hiểu đơn giản mã hóa từ là chúng ta xét từng từ, nếu từ đó chưa có trong từ điển thì thực hiện thêm vào thư viện, gán cho từ đó một con số và xem con số đó đại diện cho từ đó. Việc mã hóa từ được thực hiện với mục đích huấn luyện các mô hình LSTM về sau (do đầu vào của mô hình là một vector chứa các số).

Ở bước này, nhóm thực hiện chuyển toàn bộ chữ từ viết hoa về viết thường và sử dụng thư viện nltk để thực hiện mã hóa các từ.

Việc chuyển các từ viết hoa về viết thường cơ bản không làm thay đổi hiệu suất của mô hình sau này.

## Xử lý OOV

Có thể hiểu đơn giản là OOV xuất hiện khi bạn gặp một từ hoặc một token trong dữ liệu mà không có trong từ điển hoặc tập từ vựng đã được huấn luyện trước đó.

Ở bước này, nhóm thực hiện đếm các từ xuất hiện trong văn bản, những từ mà dưới một ngưỡng (ở đây là 2), sẽ thực hiện xem như là không có trong từ điển và sẽ gán bằng một từ đặc biệt (ở trong bài này là <unk>).

Việc chuyển các từ về unk sẽ giúp mô hình tránh việc học quá nhiều, cũng như việc xử lý các trường hợp ngoại lệ sau này. Việc xuất hiện các từ chỉ 1 lần sẽ được xem như bỏ qua, cải thiện hiệu suất mô hình.

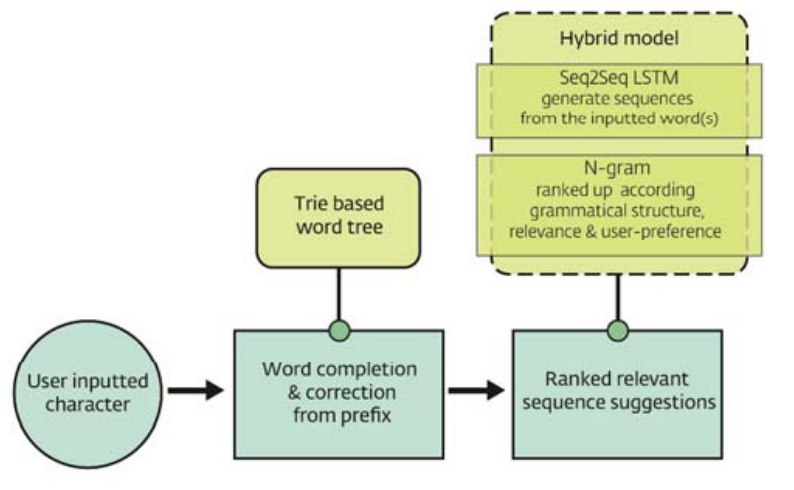
## Phân chia bộ ngữ liệu

Bộ dữ liệu sẽ được chia với tỉ lệ 8:2 tương ứng với 5772 đoạn văn bản để train và 1444 đoạn văn bản để test mô hình.

# PHƯƠNG PHÁP VÀ MÔ HÌNH

## Kiến trúc mô hình

Diagram tổng thể của mô hình:



Kiến trúc của model bao gồm 2 thành phần chính: word completion and correction from prefix và ranked relevant sequence suggestions (tạm dịch là hoàn thành và sửa lỗi và xếp hạng gợi ý trình tự liên quan).

Trong đó phần hoàn thành và sửa lỗi trong kiến trúc sẽ sử dụng cấu trúc trie (một cấu trúc đơn giản trong việc kiểm tra chính tả và dự đoán từ) và phần xếp hạng gợi ý trình tự liên quan sẽ được chia làm 2 nhiệm vụ: đưa ra từ gợi ý và xếp hạng các từ đó theo thứ tự ưu tiên. Hai nhiệm vụ này sẽ được giải quyết lần lượt bằng mô hình Sequence to sequence LSTM và mô hình NGram.

Ví dụ: ta có 3 từ “how old ares” thì mong muốn của chúng ta mô hình sẽ dự đoán ra câu “how old are you ?”. Như vậy luồng dữ liệu sẽ đi từ người dùng nhập từ “how old ares”, sau đó sẽ trải qua bước kiểm tra và sửa lỗi. Trong câu này lỗi sai nằm ở từ “ares”. Mô hình trie sẽ xử lý bước này. Tiếp đến câu được sửa lỗi là “how old are” sẽ đi qua mô hình LSTM để tìm ra từ tiếp theo là “you”. Câu đầu ra “how old are you” sẽ được trở thành câu cầu vào cho mô hình LSTM. Cứ như vậy lặp 5 lần. Sau đó ta sẽ có 5 câu với độ dài khác nhau. Sau đó, mô hình N-gram sẽ thực hiện tính toán xác suất của từng câu và lựa chọn ra câu có xác suất cao nhất. Ở đây ta mong đợi câu đầu ra sẽ là “how old are you ?”.

## Hoàn thành từ

### Cấu trúc dữ liệu Trie

Trie là một cấu trúc dữ liệu cây nâng cao, còn được gọi là cây tiền tố hoặc cây số. Trie được sử dụng để lưu trữ và tìm kiếm các chuỗi trên một bộ chữ cái hữu hạn. Trie cho phép lưu trữ hiệu quả các từ có tiền tố chung và thực hiện các truy vấn khớp mẫu một cách nhanh chóng. Trie được sử dụng trong các chương trình kiểm tra chính tả, dự đoán văn bản, tìm kiếm xấp xỉ và nhiều ứng dụng khác. Trie được mô tả lần đầu tiên bởi Axel Thue vào năm 1912 và sau đó được phát triển bởi René de la Briandais và Edward Fredkin vào những năm 1950 và 1960. Trie có thể có nhiều nhánh, mỗi nhánh đại diện cho một ký tự của chuỗi. Vị trí của một nút trong trie xác định chuỗi mà nó liên quan. Nút cuối cùng của mỗi chuỗi được đánh dấu là nút kết thúc từ. Trie có thể được cài đặt bằng nhiều ngôn ngữ lập trình khác nhau, như C++, Java, Python, C#, Javascript, v.v. Trie có nhiều ưu điểm như:

- Tìm kiếm nhanh: Trie cho phép tìm kiếm một chuỗi trong thời gian tỉ lệ với độ dài chuỗi, không phụ thuộc vào số lượng chuỗi trong tập hợp.

- Tiết kiệm bộ nhớ: Trie cho phép lưu trữ các chuỗi có tiền tố chung trong cùng một nhánh, giảm thiểu sự lặp lại và tận dụng không gian bộ nhớ.

- Duyệt theo thứ tự: Trie cho phép duyệt các chuỗi theo thứ tự từ điển, chỉ cần duyệt qua các nút từ gốc đến lá.

- Khớp mẫu hiệu quả: Trie cho phép thực hiện các truy vấn khớp mẫu, như tìm kiếm tất cả các chuỗi có tiền tố, hậu tố, hoặc chứa một chuỗi con cho trước.(1)(2)(3)(4)

Trong đồ án, trie được xây dựng dựa trên cấu trúc căn bản như trên. Khi một từ được đưa vào. Trie sẽ lần lượt xét từ đó từ node gốc đến node con. Nếu từ cuối là node lá thì xem như từ đó là đã hoàn thiện. Nếu từ node con của một chữ không chứa chữ cái tiếp theo của từ thì sẽ xem như từ đó sai chính tả và xem từ đó như “<unk>”.

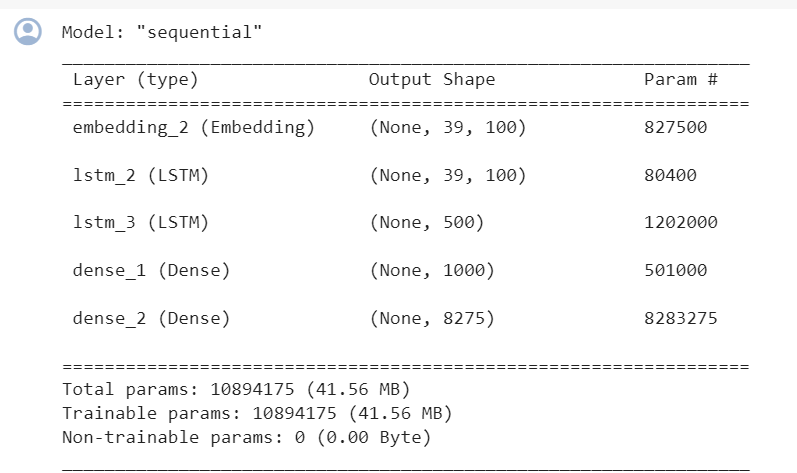
## Dự đoán câu

### Mô hình Seq2Seq LSTM

Mô hình Seq2Seq LSTM là một mô hình học sâu được sử dụng cho các bài toán xử lý ngôn ngữ tự nhiên, như dịch máy, tóm tắt văn bản, sinh văn bản, v.v. Mô hình này có hai thành phần chính là bộ mã hóa (Encoder) và bộ giải mã (Decoder), cả hai đều sử dụng mạng LSTM để xử lý các chuỗi đầu vào và đầu ra có độ dài khác nhau. Bộ mã hóa có nhiệm vụ biến đổi chuỗi đầu vào thành một vectơ trạng thái ẩn, chứa thông tin về nội dung của chuỗi. Bộ giải mã có nhiệm vụ sinh ra chuỗi đầu ra từ vectơ trạng thái ẩn, bằng cách dự đoán từng từ một theo xác suất có điều kiện. Mô hình Seq2Seq LSTM có thể ánh xạ các chuỗi có ngữ nghĩa và cấu trúc khác nhau, và có thể giải quyết các vấn đề phụ thuộc dài, ghi nhớ và biểu diễn mối quan hệ của các thông tin phụ thuộc vào ngữ cảnh trong văn bản. (5)(6)(7)(8)(9)(10)

Lý do lựa chọn mô hình LSTM: do mô hình LSTM là mô hình Seq2Seq:

Trong phạm vi đồ án môn học, do giới hạn về bộ nhớ ram, thời gian chạy GPU của google colab nên mô hình LSTM được xây dựng sẽ có cấu trúc như sau:



Đầu vào của mô hình này sẽ là một đoạn tối đa 39 từ. Các từ sẽ được đi qua các bước tiền xử lý được nêu ở phần 3. Sau đó được đưa qua mô hình để huấn luyện. Đầu ra của mô hình là một từ được dự đoán là từ tiếp theo của đoạn văn bản đó.

Trong đồ án môn học này, đầu ra mô hình sẽ tạo ra 5 từ bằng cách thêm từ cuối vào trong mô hình LSTM sau đó tiếp tục thực hiện dự đoán. Lặp 5 lần ta sẽ có 5 từ dự đoán tiếp theo của câu đầu vào.

Các kiến trúc xây dựng mô hình LSTM được tham khảo và thử nghiệm trên bộ dữ liệu được nêu ở bên trên.

## Xếp hạng các câu dự đoán

### Mô hình N-gram

Mô hình N-gram là một phương pháp thống kê để dự đoán xác suất của một từ hoặc cụm từ tiếp theo trong một văn bản, dựa trên tần suất xuất hiện của các chuỗi từ liên tiếp trong văn bản. Mô hình N-gram có thể được sử dụng trong các ứng dụng xử lý ngôn ngữ tự nhiên, như dịch máy, phân tích cảm xúc, hoàn thành từ, v.v.

Một số điểm chính của mô hình N-gram là:

- N-gram là một chuỗi n từ liền kề trong một văn bản. Ví dụ: "con bò nhảy qua bãi cỏ" là một 5-gram.

- N-gram có thể được tính bằng cách đếm số lần xuất hiện của một chuỗi n từ trong một tập dữ liệu văn bản và chia cho tổng số từ trong tập dữ liệu.

- Mô hình N-gram thường sử dụng giả định Markov, tức là xác suất của một từ chỉ phụ thuộc vào một số từ trước đó, thay vì toàn bộ câu. Ví dụ: xác suất của từ "cỏ" chỉ phụ thuộc vào từ "bãi" và "qua" trong một mô hình 3-gram.

- Mô hình N-gram có thể gặp vấn đề khi gặp các chuỗi từ chưa từng xuất hiện trong tập dữ liệu, hoặc có xác suất quá nhỏ. Để khắc phục vấn đề này, người ta sử dụng các phương pháp làm mịn, như cộng một, giảm trừ, back-off, hay nội suy, để phân bổ lại xác suất cho các chuỗi từ. (11)(12)(13)(14)(15)(16)

Mô hình N-Gram trong đề tài này được sử dụng để đánh giá xếp hạng xem trong các câu được tạo ra bởi mô hình LSTM thì câu nào có xác suất xuất hiện cao nhất. Từ đó sẽ thực hiện xếp hạng các câu đề xuất dựa trên xác suất cao thấp. Để tránh xác suất bằng 0 thì nhóm có thực hiện việc smoth cho mô hình N-Gram này. Do các câu có độ dài khác nhau sẽ ảnh hưởng đến xác suất của các câu, nên các câu ngắn sẽ được thêm vào đuôi từ “<unk>” để so sánh với các câu dài hơn.

# ĐÁNH GIÁ

Để thuận tiện đánh giá mô hình, nhóm sẽ thực hiện đầu vào gồm 4 từ đầu tiên trong tập test. Kết quả dự đoán sẽ là từ có xác suất ngram cao nhất. Dựa vào độ dài của dự đoán để so sánh với câu trong tập test. Nhóm sẽ thực hiện đánh giá dựa trên 3 thang đo BLEU-score, ROUGE-score, BERT-score

## Thang đo BLEU

Thang đo BLEU, viết tắt của Bilingual Evaluation Understudy, là một phương pháp đánh giá chất lượng bản dịch tự động dựa trên các bản dịch tham khảo. BLEU được thiết kế để sử dụng trong dịch máy, nhưng cũng được áp dụng trong các nhiệm vụ khác như tóm tắt văn bản, nhận dạng giọng nói, sinh nhãn ảnh, v.v. BLEU đánh giá một câu thông qua việc so khớp câu đó với các câu mẫu và cho thang điểm từ 0 (sai lệch tuyệt đối) đến 1 (khớp tuyệt đối). BLEU được biết đến như một phương pháp đơn giản, dễ hiểu, chi phí tính toán thấp và tương đồng với cách đánh giá của con người.

Cách tính của BLEU là đếm số n-gram khớp nhau giữa câu mẫu (R) và câu được đánh giá (C) sau đó chia cho số token của C. Việc chọn n phụ thuộc vào ngôn ngữ, nhiệm vụ và mục tiêu cụ thể. Đơn giản nhất ta có thể sử dụng unigram là n-gram chứa 1 token ( n=1 ). Một cách trực quan, n càng lớn, câu văn càng mượt.

Tuy nhiên, BLEU cũng có nhược điểm là không thể đánh giá được thứ tự của từ. Đây vừa là ưu điểm vừa là hạn chế của BLEU. Trong ngôn ngữ, một câu có thể được biểu diễn bởi các thứ tự từ khác nhau nhưng vẫn phải tuân theo những quy tắc nhất định. (17)(18)(19)(20)

Kết quả trung bình khi tính BLEU-score trên tập test:

## Thang đo ROUGE

Thang đo ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một phương pháp đánh giá chất lượng của các hệ thống tóm tắt văn bản tự động. Nó hoạt động bằng cách so sánh văn bản tóm tắt tự động với văn bản tóm tắt tham chiếu, thường được tạo bởi con người.

Có nhiều biến thể của ROUGE, bao gồm:

- ROUGE-N: Đo số lượng n-gram phù hợp giữa văn bản do mô hình tạo ra và tham chiếu do con người tạo ra.

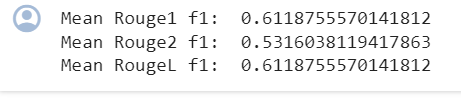
- ROUGE-L: Dựa trên chuỗi con chung dài nhất (LCS) giữa đầu ra mô hình và tham chiếu của chúng tôi, tức là chuỗi từ dài nhất (không nhất thiết phải liên tiếp nhưng vẫn theo thứ tự) được chia sẻ giữa cả hai.

- ROUGE-S: Là một phương pháp cho phép chúng ta áp dụng một mức độ khoan dung cho việc khớp n-gram được thực hiện bằng ROUGE-N và ROUGE-L. ROUGE-S là một độ đo đồng xuất hiện skip-gram: điều này cho phép tìm kiếm các từ liền kề từ văn bản tham chiếu xuất hiện trong đầu ra của mô hình nhưng được phân cách bởi một hoặc nhiều từ khác.

Mỗi biến thể của ROUGE có ưu và nhược điểm riêng, và lựa chọn phù hợp sẽ phụ thuộc vào nhiệm vụ cụ thể. (21)(22)(23)(24)

Nhóm sẽ tính f1-score rouge trên 3 thang đo ROUGE1, ROUGE2, ROUGEL

Kết quả trung bình khi tính ROUGE-score trên tập test:



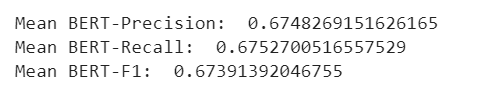
## Thang đo BERT

BERT (Bidirectional Encoder Representations from Transformers) là một kỹ thuật học máy dựa trên các transformer được dùng cho việc huấn luyện trước xử lý ngôn ngữ tự nhiên (NLP) được phát triển bởi Google. BERT giúp tối ưu hóa hệ thống tìm kiếm, hiểu được gần chính xác sắc thái, ý nghĩa ngữ cảnh trong câu.

Ví dụ, nếu tìm kiếm cụm từ "cow fishing", thật chất đây là tên một loài cá (cá nược). Tuy nhiên khi bạn gõ cụm từ về "cow fishing", Google sẽ cung cấp kết quả liên quan đến chăn nuôi bò. Mặc dù bạn đã cố tình sử dụng từ "fishing" để cung cấp ngữ cảnh, Google đã bỏ qua bối cảnh đó và cung cấp kết quả liên quan đến bò.

BERT ra đời để giải quyết điều này, nó giúp phân biệt những sắc thái nghĩa trừu tượng nhất, đưa ra những kết quả phù hợp và có liên quan nhất. (25)(26)(27)

Kết quả tính BERT-score trên tập test:



Đánh giá kết quả thu được qua 3 thang đo, có thể thấy được rằng kết quả thu được không quá tốt. So với bài báo (\*) được tham khảo (81% - 84%) thì vẫn còn thua kém nhiều.

Kết quả thu được so với bài báo có sự chênh lệch do ngữ liệu của 2 bài là khác nhau, chữ Bangla là chữ tượng hình, kèm theo đó là sự đa dạng của các đoạn văn twitter. Tiếp theo đó là do giới hạn về dung lượng do google colab cung cấp nên chưa thể huấn luyện trên bộ dữ liệu lớn hơn. Cuối cùng là về việc xây dựng mô hình LSTM chưa thật sự tối ưu khiến cho việc dự đoán chưa đủ tốt.

# TÀI LIỆU THAM KHẢO

(\*) Bài báo tham khảo: Sci-Hub | Word Completion and Sequence Prediction in Bangla Language Using Trie and a Hybrid Approach of Sequential LSTM and N-gram. 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT) | 10.1109/icaict51780.2020.9333518  
<https://sci-hub.se/https://ieeexplore.ieee.org/abstract/document/9333518>

(1) Trie - Wikipedia. <https://en.wikipedia.org/wiki/Trie>.

(2) Trie | (Insert and Search) - GeeksforGeeks. <https://www.geeksforgeeks.org/trie-insert-and-search/>.

(3) Trie Data Structure - Explained with Examples - Studytonight. <https://www.studytonight.com/advanced-data-structures/trie-data-structure-explained-with-examples>.

(4) Tries or Prefix Trees | Baeldung on Computer Science. <https://www.baeldung.com/cs/tries-prefix-trees>.

(5) Giới thiệu bài toán Tóm tắt Văn bản sử dụng mô hình LSTM-based Seq2seq .... <https://viblo.asia/p/gioi-thieu-bai-toan-tom-tat-van-ban-su-dung-mo-hinh-lstm-based-seq2seq-va-co-che-attention-6J3ZgWeRZmB>.

(6) Seq2Seq - Hiểu về mô hình Encoder-Decoder - Trí tuệ nhân tạo. <https://trituenhantao.io/kien-thuc/hieu-ve-mo-hinh-encoder-decoder-seq2seq/>.

(7) Giải thích chi tiết về mô hình Sequence-to-Sequence. <https://bing.com/search?q=M%c3%b4+h%c3%acnh+Seq2Seq+LSTM>.

(8) Giải thích chi tiết về mô hình Sequence-to-Sequence. <https://nguyentruonglong.net/giai-thich-chi-tiet-ve-mo-hinh-sequence-to-sequence.html>.

(9) NLP [P3] - Seq2Seq Model - Hạt nhân của Google Translate. <https://nerophung.github.io/2019/05/18/nlp-p3>.

(10) ỨNG DỤNG SEQ2SEQ-LSTM TRONG MÔ HÌNH DỰ BÁO NGẮN HẠN PHỤ TẢI CHO LƯỚI .... <http://jst.tnu.edu.vn/jst/article/view/9060>.

(11) Một số phương pháp làm mịn trong mô hình trong mô hình N-gram. <https://viblo.asia/p/mot-so-phuong-phap-lam-min-trong-mo-hinh-trong-mo-hinh-n-gram-aqkRnbMMRnA>.

(12) Hướng dẫn đầy đủ về mô hình ngôn ngữ N-Gram trong Xử lý ngôn ngữ tự nhiên. <https://ichi.pro/vi/huong-dan-day-du-ve-mo-hinh-ngon-ngu-n-gram-trong-xu-ly-ngon-ngu-tu-nhien-229134357142740>.

(13) N-gram – Wikipedia tiếng Việt. <https://vi.wikipedia.org/wiki/N-gram>.

(14) Mô hình N-grams - Thêm dấu tiếng Việt | VN AIDr - GitHub Pages.

<https://bing.com/search?q=M%c3%b4+h%c3%acnh+N-gram>.

(15) n-gram - Wikipedia. <https://en.wikipedia.org/wiki/N-gram>.

(16) Giải thích n-gram là gì và ứng dụng trong xử lý ngôn ngữ tự nhiên. <https://xaydungso.vn/blog/giai-thich-n-gram-la-gi-va-ung-dung-trong-xu-ly-ngon-ngu-tu-nhien-vi-cb.html>.

(17) BLEU - Phép đo trong dịch máy - Trí tuệ nhân tạo. <https://trituenhantao.io/kien-thuc/bleu-phep-do-trong-dich-may/>.

(18) Phân biệt các loại thang đo trong nghiên cứu - Hệ thống thông tin Thống .... <https://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/720-phan-biet-thang-do-trong-nghien-cuu>.

(19) Tìm hiểu về độ đo BLEU | <Vunb />. <https://vunb.github.io/post/2019/07/tim-hieu-bleu/>.

(20) Độ trắng – Wikipedia tiếng Việt. <https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99_tr%E1%BA%AFng>.

(21) 2 phút để hiểu cách tính ROUGE metric - Viblo. <https://viblo.asia/p/2-phut-de-hieu-cach-tinh-rouge-metric-5OXLAPEBJGr>.

(22) Thang Đo Trong Spss: Khái Niệm, 4 Loại Và Ví Dụ. <https://trithuccongdong.net/tai-lieu-spss/thang-do-trong-spss.html>.

(23) 2 phút để hiểu cách tính ROUGE metric - Viblo. <https://bing.com/search?q=Thang+%c4%91o+ROUGE>.

(24) Nhóm sinh viên UIT có bài báo được đăng tại một trong những tạp chí .... <https://tuyensinh.uit.edu.vn/nhom-sinh-vien-uit-co-bai-bao-duoc-dang-tai-mot-trong-nhung-tap-chi-thuoc-xep-hang-cao-nhat-scie-q1>.

(25) Bert là gì và Bert ảnh hưởng thế nào tới ... - MarketingAI. <https://marketingai.vn/bert-la-gi-va-bert-anh-huong-the-nao-toi-cong-cu-tim-kiem-tren-google-19493740.htm>.

(26) Berg Balance Scale - PHCN Online. <https://phcn-online.com/wp-content/uploads/2021/04/berg_balance_scale_vn.pdf>.

(27) BERT (mô hình ngôn ngữ) – Wikipedia tiếng Việt. <https://vi.wikipedia.org/wiki/BERT_%28m%C3%B4_h%C3%ACnh_ng%C3%B4n_ng%E1%BB%AF%29>.