

# ChatGPT: Bắt đầu Kết thúc Chú thích Dữ liệu Ngôn ngữ Thủ công? Trường hợp sử dụng "Nhận dạng thể loại tự động"

GVHD: Mr. Nguyễn Hồng Bửu Long  
Mr. Lương An Vinh

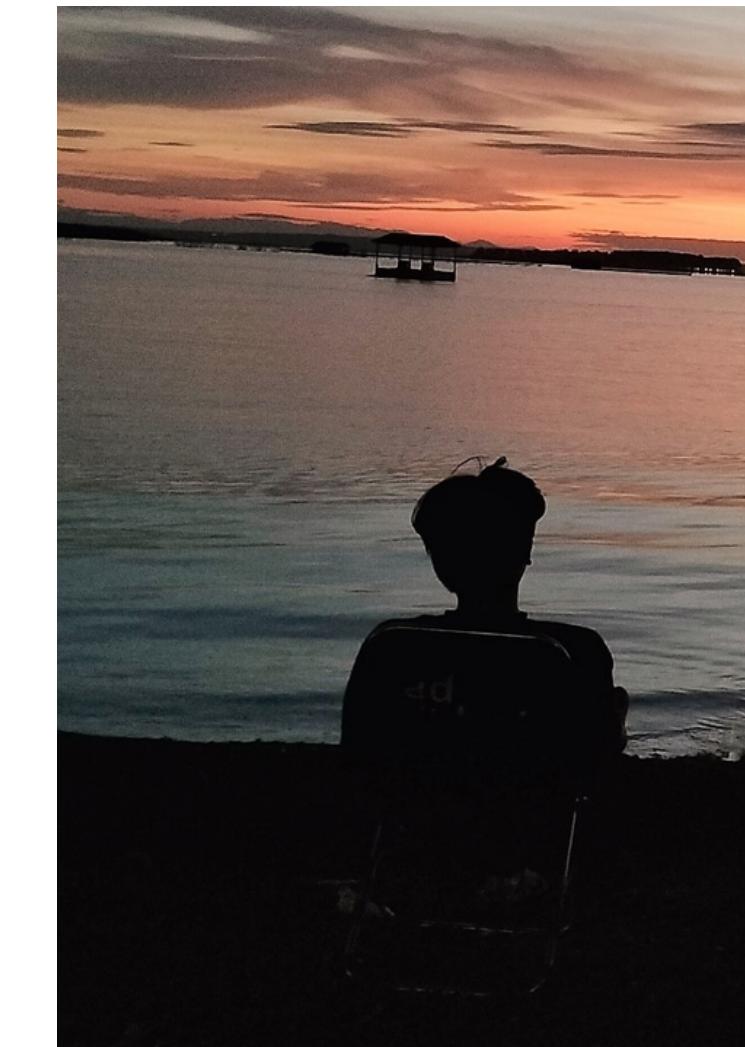


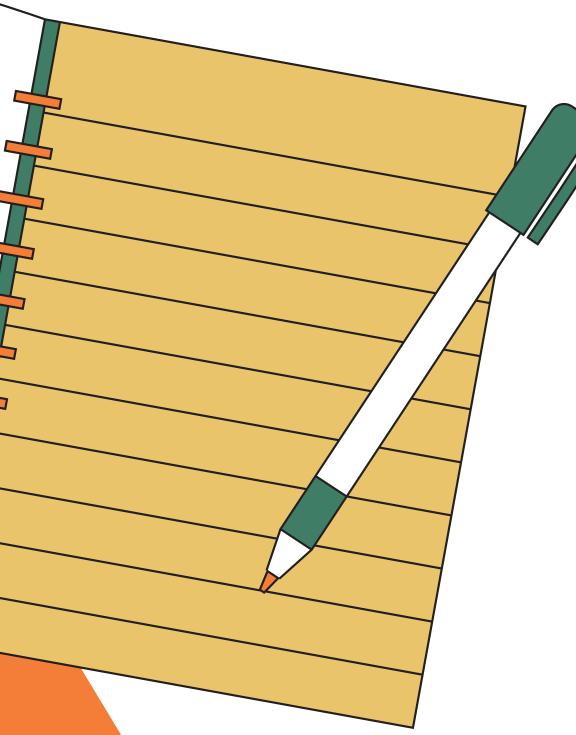
# \*\* CÁC THÀNH VIÊN TRONG NHÓM:

LÊ ĐỨC ĐẠT - 20127674



BÙI TUÂN DUNG - 20127141





# 1. Tóm tắt sơ lược bài báo:



**Tác giả nói về việc sử dụng ChatGPT, một mô hình ngôn ngữ tự động tạo văn bản, cho việc phân loại văn bản theo thể loại (genre) mà không cần đào tạo (zero-shot).**

**Phân loại văn bản theo thể loại (genre)** là quá trình gán nhãn hoặc xác định thể loại của một văn bản, dựa trên nội dung, cấu trúc, ngôn ngữ và đặc điểm khác của văn bản đó. Thể loại có thể bao gồm tiểu thuyết, truyện ngắn, kịch, thơ, báo cáo khoa học, bài báo, văn bản luật pháp và nhiều thể loại khác.

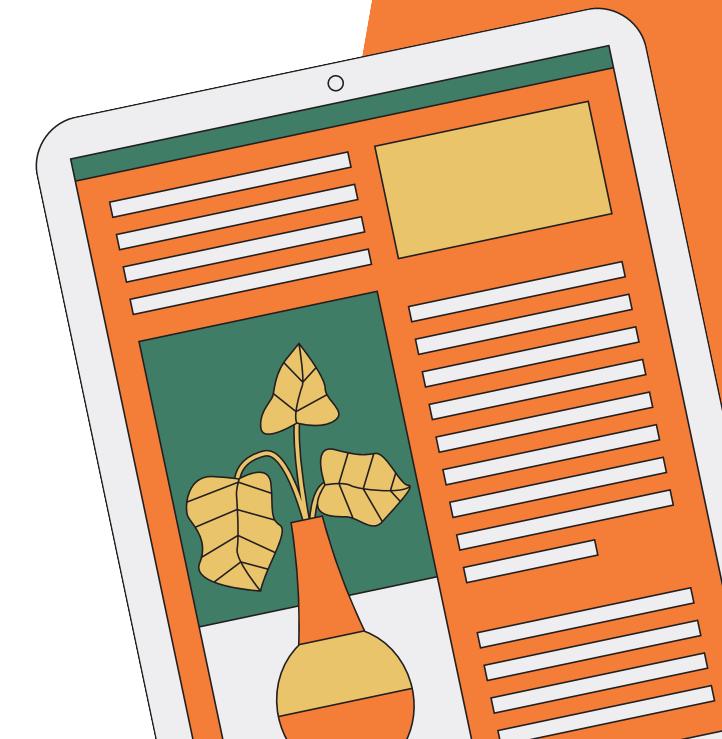
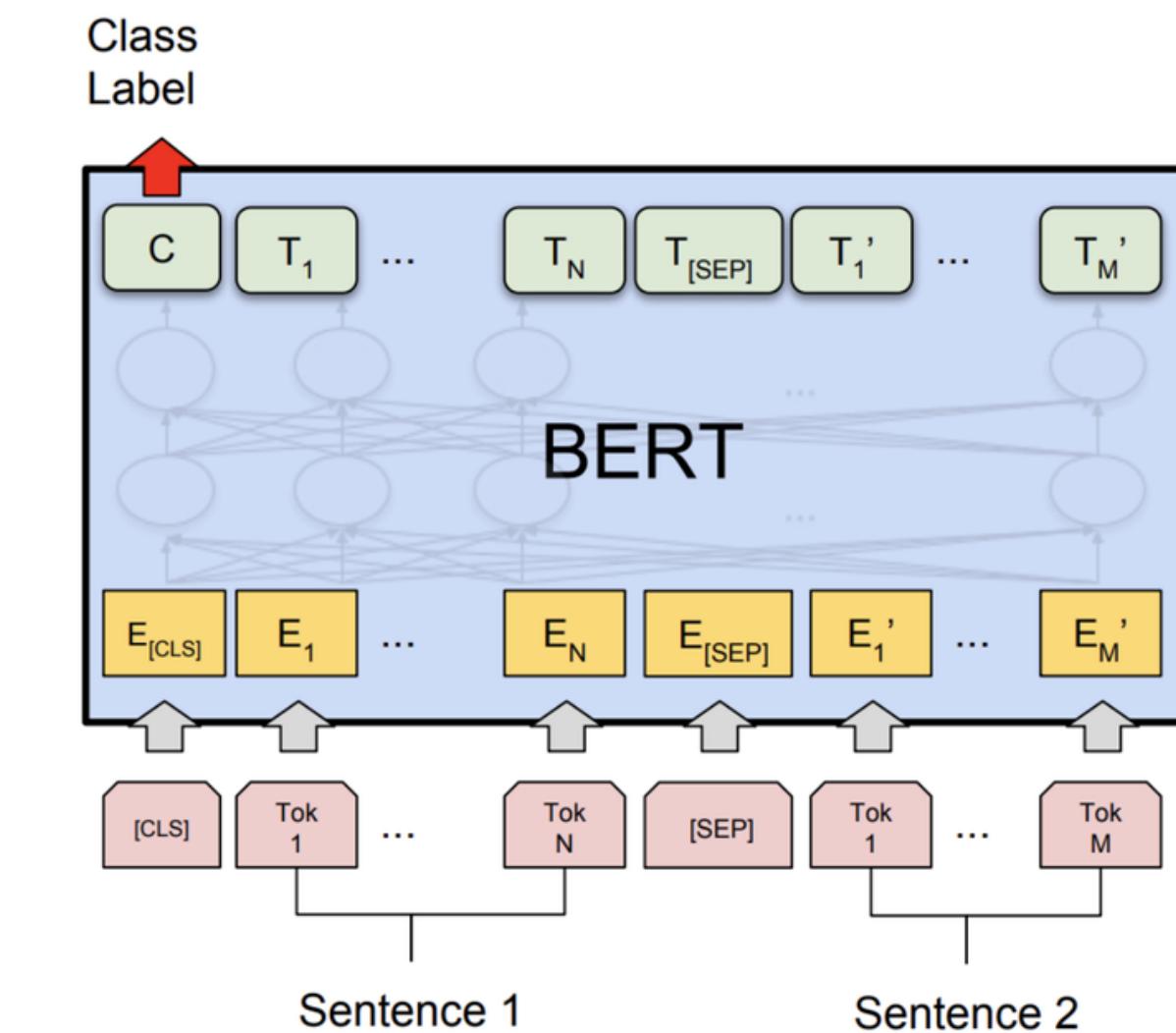
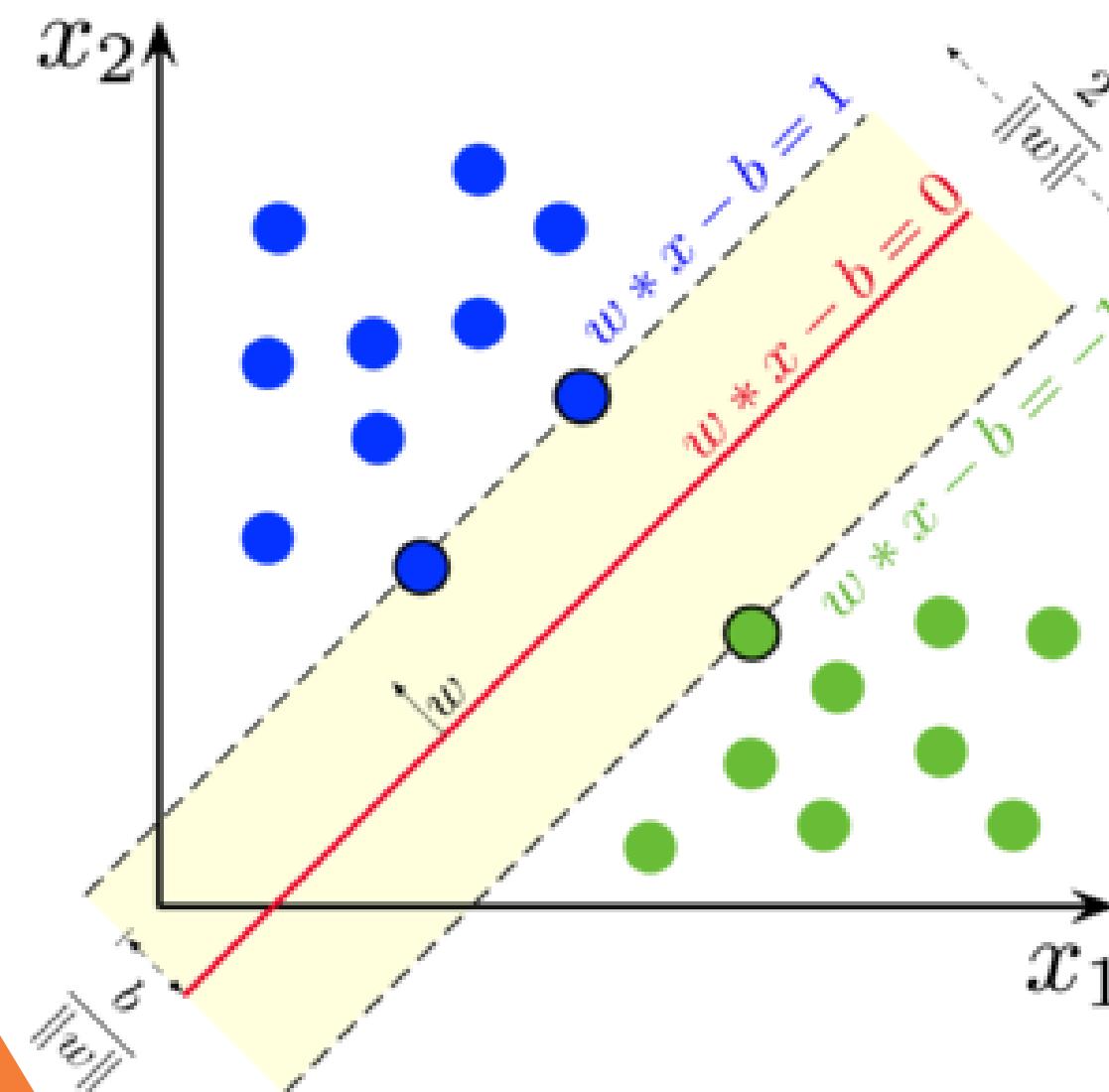
Phân loại văn bản theo thể loại thường được sử dụng trong các ứng dụng như phân loại sách, phân tích dữ liệu văn bản, hệ thống khuyến nghị nội dung, và cải thiện trải nghiệm người dùng trên các nền tảng đọc sách điện tử hoặc tìm kiếm thông tin.

Phân loại văn bản theo thể loại thường được sử dụng trong các ứng dụng như phân loại sách, phân tích dữ liệu văn bản, hệ thống khuyến nghị nội dung, và cải thiện trải nghiệm người dùng trên các nền tảng đọc sách điện tử hoặc tìm kiếm thông tin.





Các phương pháp phân loại văn bản theo thể loại thường dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy, bao gồm cả học có giám sát và học không giám sát. Các thuật toán phổ biến cho việc phân loại văn bản bao gồm Naive Bayes, hỗ trợ máy vectơ (SVM), mạng nơ-ron nhân tạo, và các mô hình dựa trên biểu diễn ngôn ngữ như BERT, GPT và các mô hình tương tự.

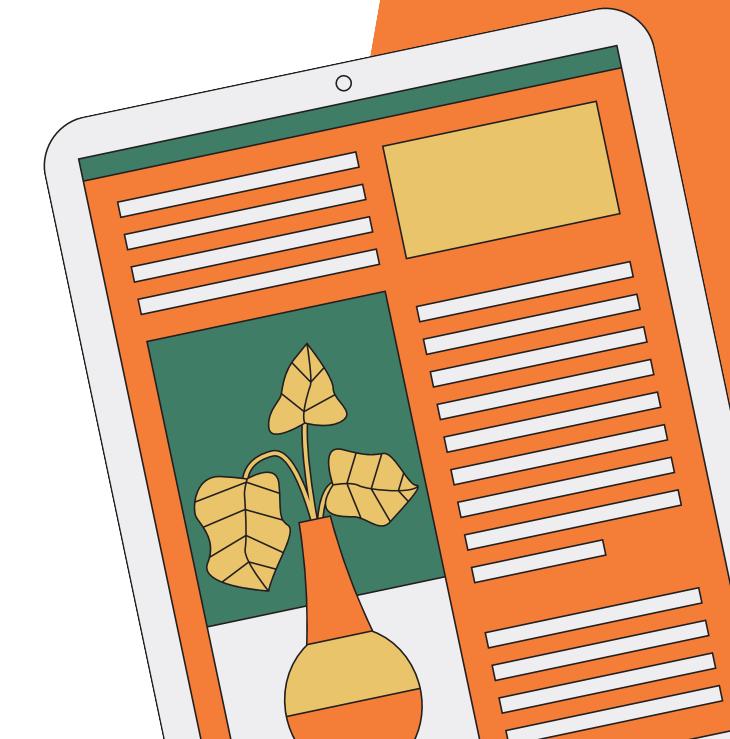
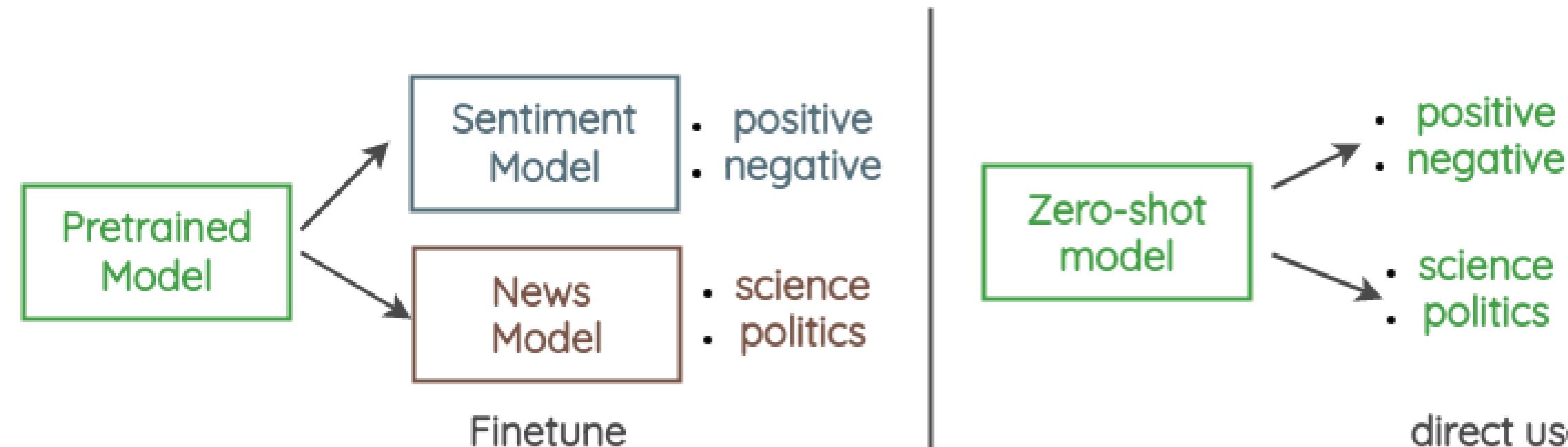




**Zero-shot text classification (phân loại văn bản không cần dữ liệu huấn luyện)** là một phương pháp phân loại văn bản mà không cần dữ liệu huấn luyện có nhãn cho từng thể loại cụ thể. Thay vào đó, mô hình sử dụng kiến thức đã học từ dữ liệu huấn luyện tổng quát (đã được tiền huấn luyện) để phân loại văn bản dựa trên ngữ cảnh và đặc điểm ngôn ngữ của chúng.

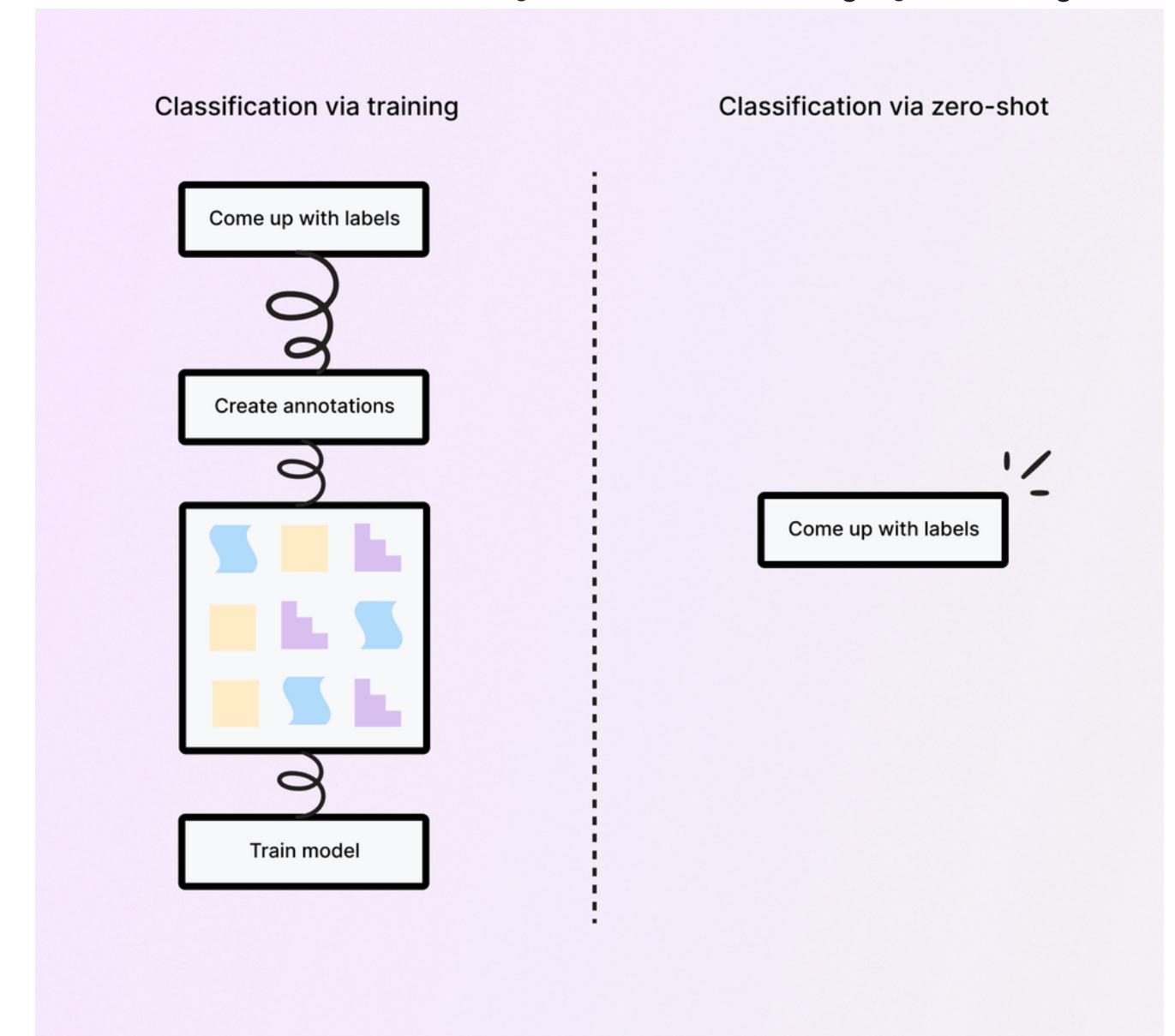
Các mô hình zero-shot thường dựa trên các mô hình học sâu như BERT, GPT, RoBERTa, và các kiến trúc tương tự. Các mô hình này được tiền huấn luyện trên một lượng lớn dữ liệu văn bản không nhãn từ nhiều nguồn khác nhau, giúp chúng học được các đặc trưng ngôn ngữ tổng quát và ngữ nghĩa của văn bản.

### Transfer Learning vs Zero-Shot Learning

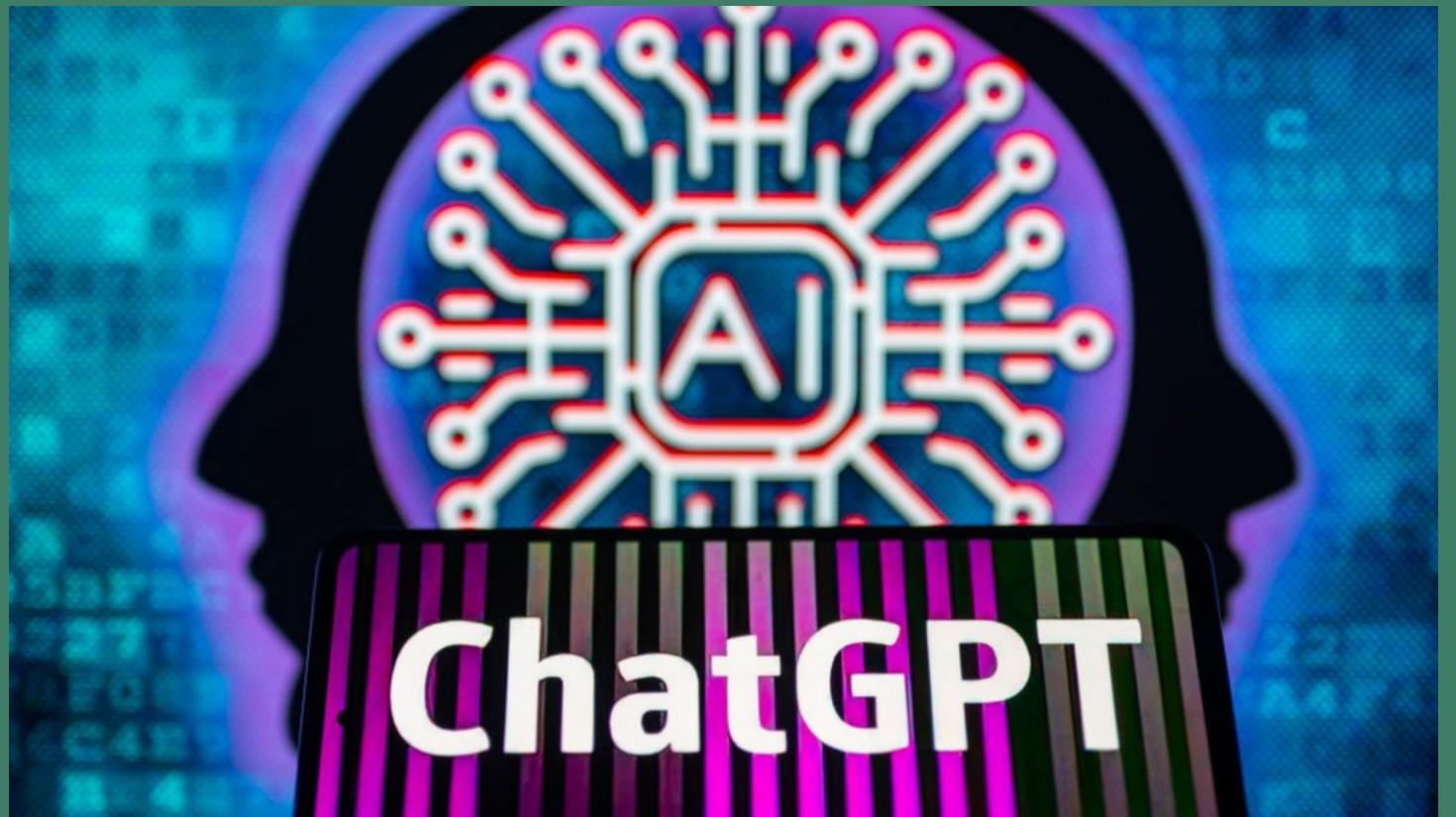


Khi áp dụng cho bài toán phân loại văn bản, mô hình zero-shot có thể nhận diện và phân loại các thể loại mà nó chưa từng gặp trong quá trình huấn luyện. Điều này làm cho zero-shot text classification đặc biệt hữu ích trong các tình huống có ít hoặc không có dữ liệu huấn luyện có nhãn, hoặc khi cần phân loại văn bản theo các thể loại mới mà mô hình chưa từng gặp.

Tuy nhiên, zero-shot text classification có thể không hoạt động tốt như các phương pháp học có giám sát đối với các bài toán phân loại văn bản có dữ liệu huấn luyện đầy đủ và cụ thể.



Bài báo so sánh ChatGPT với một mô hình ngôn ngữ khác đã được đào tạo, ở đây là một mô hình ngôn ngữ XLM-RoBERTa đa ngôn ngữ đã được tinh chỉnh trên bộ dữ liệu, được chú thích theo cách thủ công với các thể loại để phân loại văn bản theo thể loại trên các tập dữ liệu tiếng Anh và Slovenian. Kết quả cho thấy ChatGPT hoạt động tốt hơn so với mô hình đã được đào tạo trên tập dữ liệu tiếng Anh và đạt kết quả tương đương trên tập dữ liệu tiếng Slovenian.



Tuy nhiên, khi ChatGPT được sử dụng trên tiếng Slovenian mà không có dữ liệu đào tạo, kết quả không tốt bằng khi được sử dụng trên tiếng Anh, cho thấy giới hạn của ChatGPT trong việc sử dụng trên các ngôn ngữ nhỏ hơn. Bài báo cho rằng kết quả nghiên cứu này mở ra triển vọng cho việc không cần phải tốn công sức để đào tạo mô hình phân loại thể loại văn bản trên các ngôn ngữ nhỏ hơn, chẳng hạn như tiếng Slovenian.

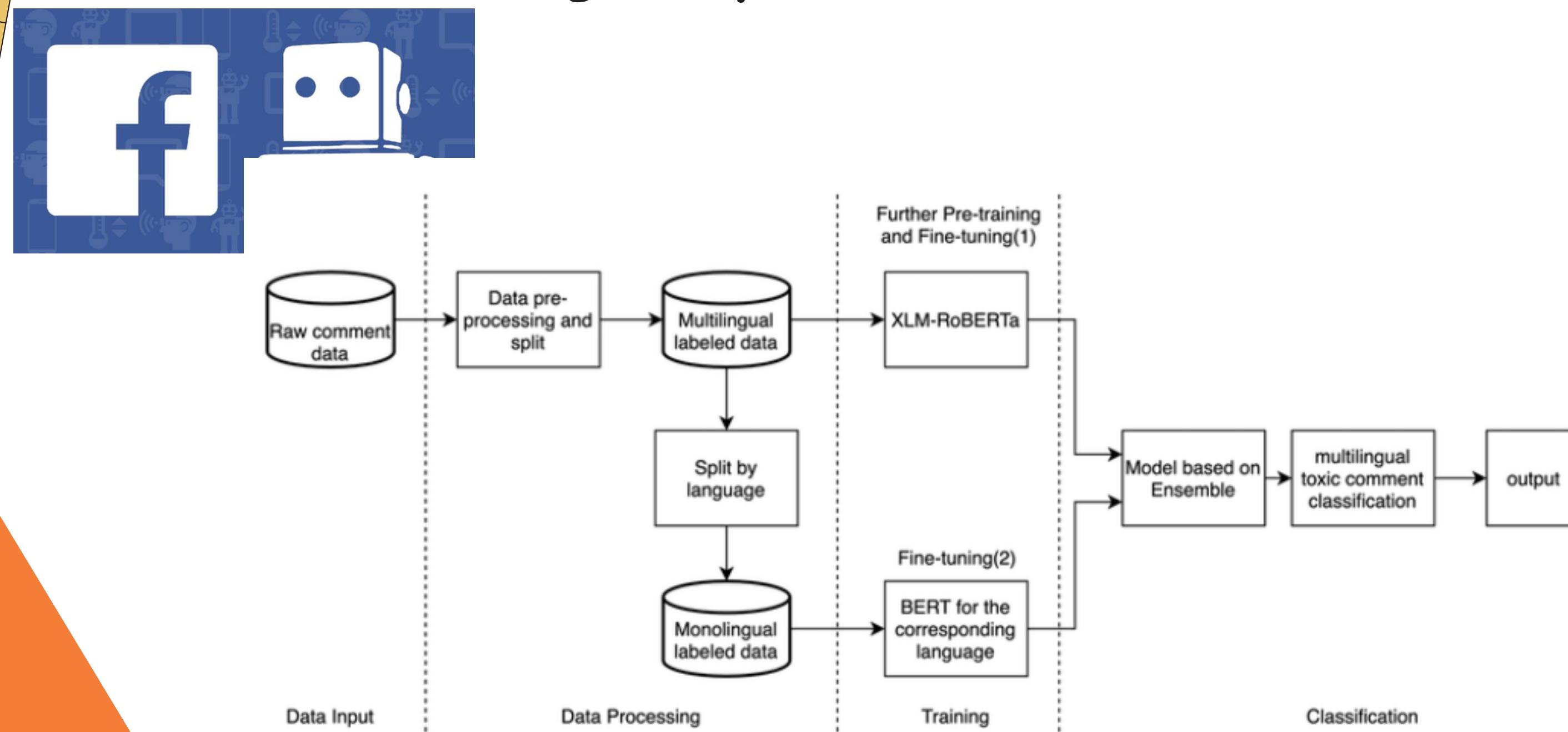
CH Hej TI, MED RONALDO IN MESSI, KDO JE BOLJŠI?

 Hej! Vprašanje o tem, ali je Ronaldo ali Messi boljši, je že dolgo predmet razprave med nogometnimi navdušenci. Oba igralca sta izjemno nadarjena in sta dosegla številne uspehe v svojih karierah.

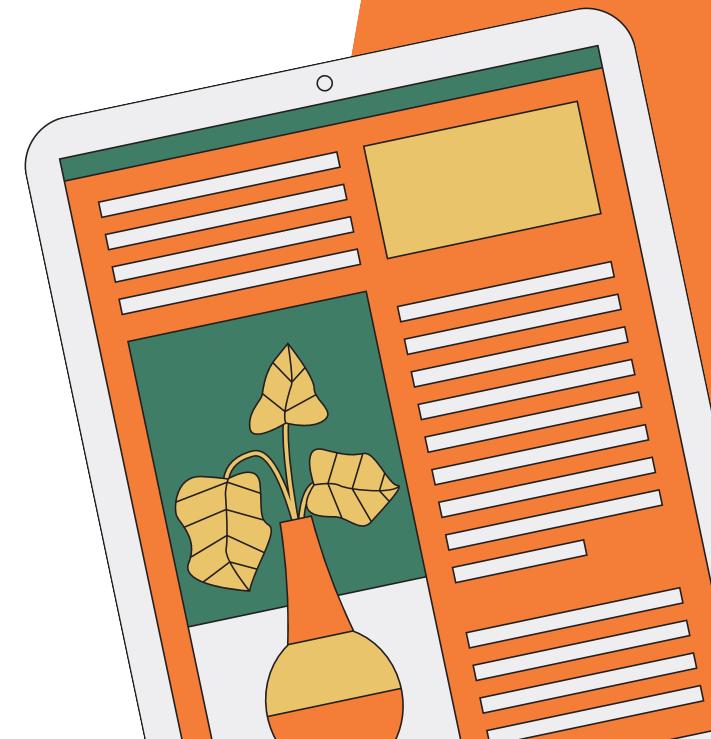
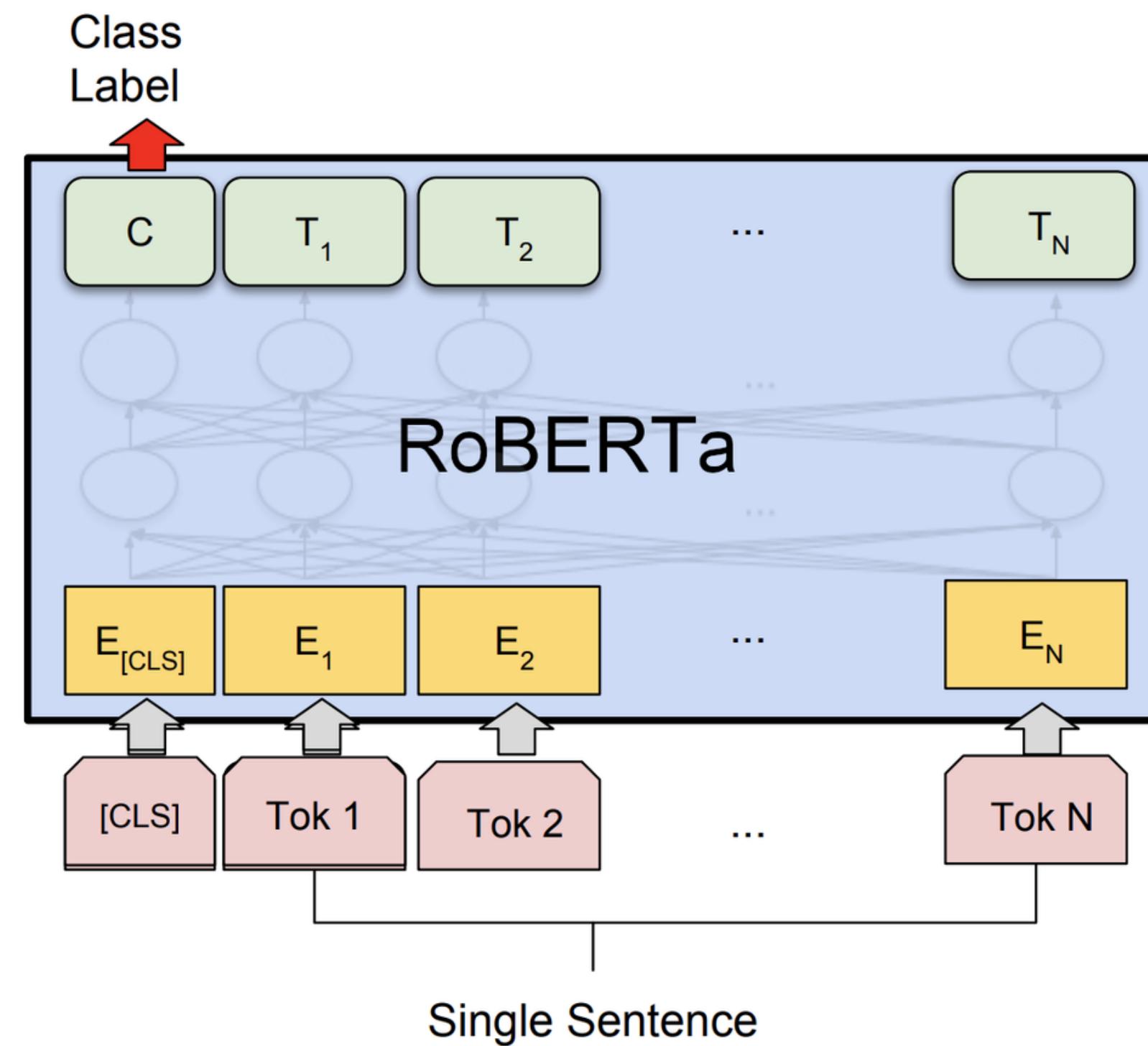
Cristiano Ronaldo je znan po svoji neverjetni atletski sposobnosti, moči in hitrosti ter izjemni sposobnosti doseganja golov. Po drugi strani pa je Lionel Messi znan po svoji izjemni dribbling tehnikah, kreativnosti na igrišču in naravnem talentu.

Kateri igralec je boljši, je stvar osebnega mnenja in preference. Nekateri ljudje raje imajo Ronaldo zaradi njegovega fizičnega sloga igre, medtem ko drugi menijo, da je Messi boljši zaradi njegove tehnične spremnosti. Oba igralca sta nesporno med najboljšimi v zgodovini nogometa, zato je težko izbrati enega kot absolutnega zmagovalca.

Nói thêm về mô hình ngôn ngữ XLM-RoBERTa, (Cross-lingual Language Model - RoBERTa) là một mô hình học sâu dựa trên kiến trúc Transformer, được thiết kế để xử lý các tác vụ xử lý ngôn ngữ tự nhiên (NLP) đa ngôn ngữ. Mô hình này được phát triển bởi Facebook AI và kết hợp hai phương pháp tiên tiến trong lĩnh vực NLP: RoBERTa và XLM.



+ RoBERTa (Robustly optimized BERT pretraining approach) là một phiên bản tối ưu hóa của mô hình BERT, được phát triển bởi nhóm nghiên cứu của Facebook AI. RoBERTa tập trung vào việc cải thiện hiệu suất của BERT thông qua việc điều chỉnh quy trình tiền huấn luyện, bao gồm việc tăng kích thước dữ liệu huấn luyện, tăng tốc độ học và loại bỏ phần "Next Sentence Prediction" trong kiến trúc BERT.



+ XLM (Cross-lingual Language Model) là một mô hình dựa trên Transformer được thiết kế để xử lý các tác vụ NLP đa ngôn ngữ, phát triển bởi Facebook AI. XLM sử dụng kỹ thuật "unsupervised machine translation" để học biểu diễn đa ngôn ngữ thông qua việc tiền huấn luyện trên dữ liệu song ngữ. Dùng 1 trong 3 mô hình sau:

- Mô hình hóa ngôn ngữ nhân quả - lập mô hình xác suất của một từ được cung cấp cho các từ trước đó trong một câu.
- Mô hình hóa ngôn ngữ ẩn - mục tiêu mô hình hóa ngôn ngữ ẩn của BERT.
- Mô hình hóa ngôn ngữ dịch thuật - một mục tiêu mô hình hóa ngôn ngữ dịch thuật (mới) để cải thiện việc đào tạo trước đa ngôn ngữ.



XLM-RoBERTa kết hợp ưu điểm của cả RoBERTa và XLM để tạo ra một mô hình đa ngôn ngữ hiệu quả. XLM-RoBERTa được tiên huấn luyện trên một bộ dữ liệu đa ngôn ngữ lớn (100 ngôn ngữ) được gọi là CommonCrawl. Kết quả là một mô hình có khả năng xử lý và hiểu nhiều ngôn ngữ khác nhau một cách hiệu quả, đồng thời đạt được hiệu suất cao trong nhiều tác vụ NLP đa ngôn ngữ như phân loại văn bản, dịch máy, và hỏi đáp.

Tóm lại, XLM-RoBERTa là một mô hình NLP đa ngôn ngữ mạnh mẽ, kế thừa các ưu điểm của RoBERTa và XLM, và được huấn luyện trên một lượng lớn dữ liệu đa ngôn ngữ để đạt được hiệu suất cao trong các tác vụ NLP đa ngôn ngữ.



## 2. Giới thiệu:



ChatGPT đã cho thấy khả năng mạnh mẽ trong hệ thống đồi thoại, cung cấp câu trả lời rõ ràng và hữu ích hơn so với con người. Tuy nhiên, hiệu suất của nó trong các tác vụ phân loại văn bản, cụ thể là xác định thể loại tự động, chưa được so sánh với các mô hình ngôn ngữ lớn như XLM-RoBERTa. Mặc dù ChatGPT mới được công bố cách đây vài tháng, nhưng đã có nhiều nghiên cứu phân tích tiềm năng của mô hình này trong các tác vụ NLP. Kết quả cho thấy ChatGPT thường kém hơn so với các mô hình ngôn ngữ lớn tinh chỉnh.

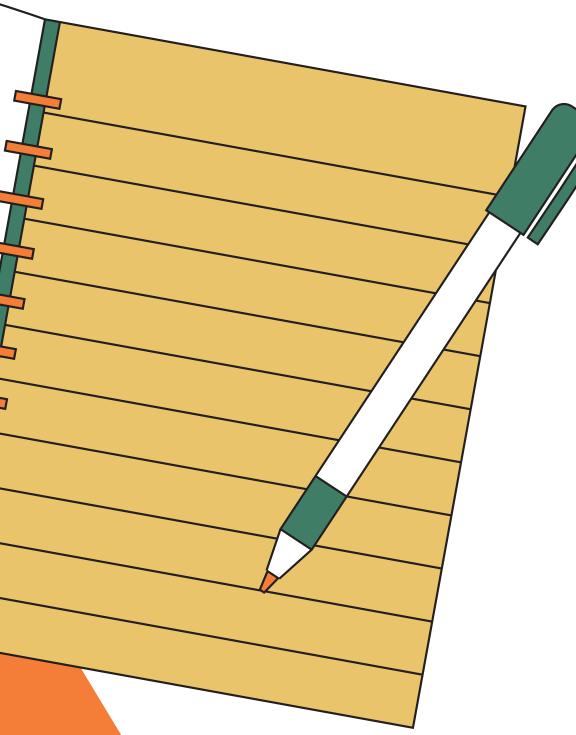
AN Quang Trung và Nguyễn Huệ có phải là hai anh em không?

 Không, Quang Trung và Nguyễn Huệ không phải là hai anh em. Quang Trung (1732-1792) là một vị tướng và nhà lãnh đạo quân sự nổi tiếng trong lịch sử Việt Nam, còn Nguyễn Huệ (1753-1792) cũng là một vị tướng và nhà lãnh đạo quân sự có công trong việc đánh bại quân Nguyễn, lên ngôi vua với tên là Quang Trung sau đó. Tuy nhiên, Quang Trung và Nguyễn Huệ không có quan hệ họ hàng hay anh em ruột với nhau.

AN Nguyễn Ánh và Gia Long có phải là hai anh em không?

 Không, Nguyễn Ánh và Gia Long không phải là hai anh em mà là cùng một người. Nguyễn Ánh là người sáng lập và đặt nền móng cho triều đại nhà Nguyễn, trong khi Gia Long là hoàng đế đầu tiên của triều đại này.



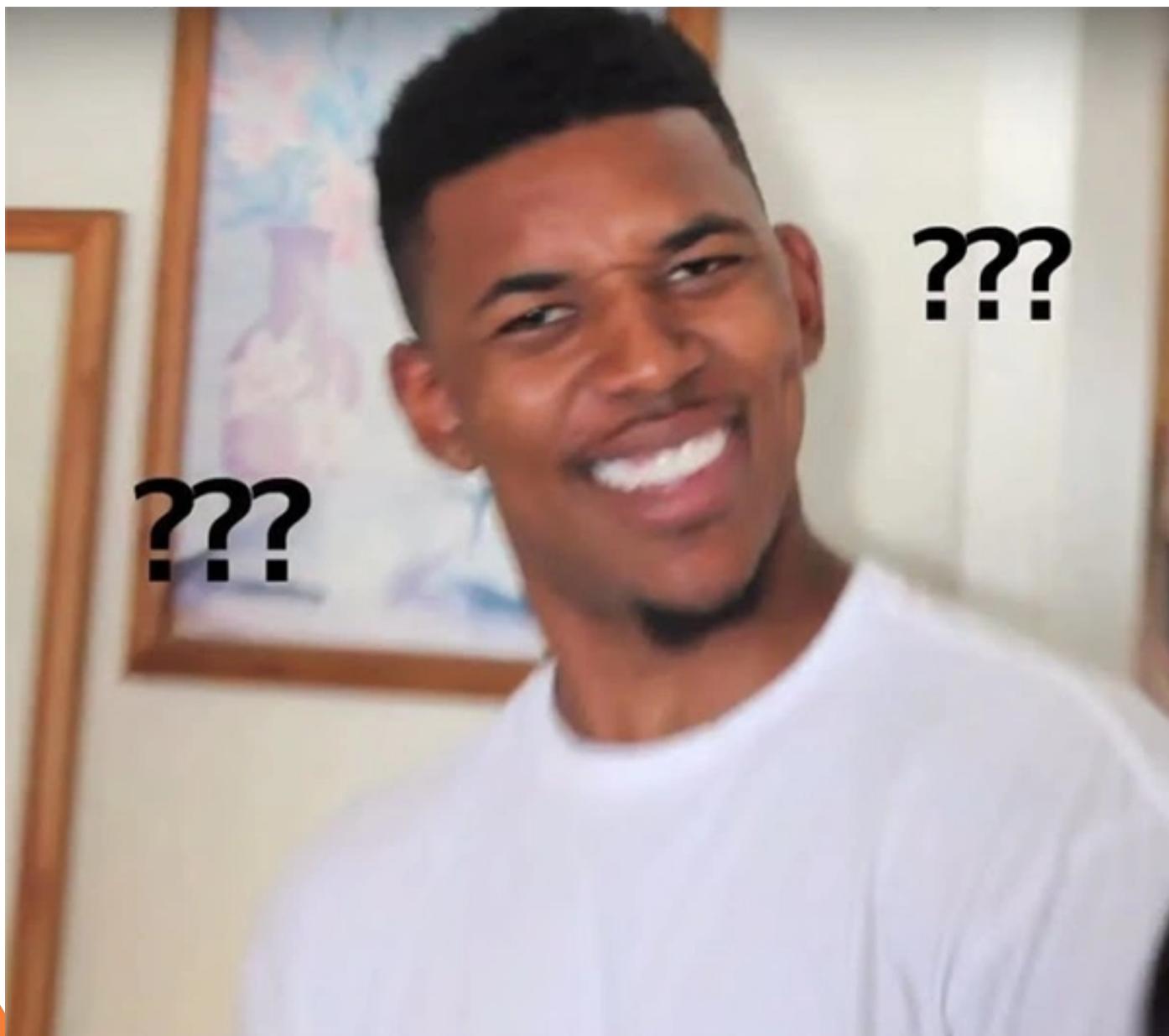


Trong một số trường hợp, ChatGPT đạt kết quả tốt hơn như trong tác vụ phát hiện quan điểm. ChatGPT cũng đạt hiệu suất cạnh tranh trong dịch máy cho các ngôn ngữ có nguồn lực cao, nhưng kém hơn đối với các ngôn ngữ có nguồn lực thấp. Một số nghiên cứu cũng so sánh hiệu suất của mô hình với chủ thích thủ công, cho thấy tiềm năng của ChatGPT trong việc phát hiện và giải thích ngôn ngữ thù địch tiềm ẩn.



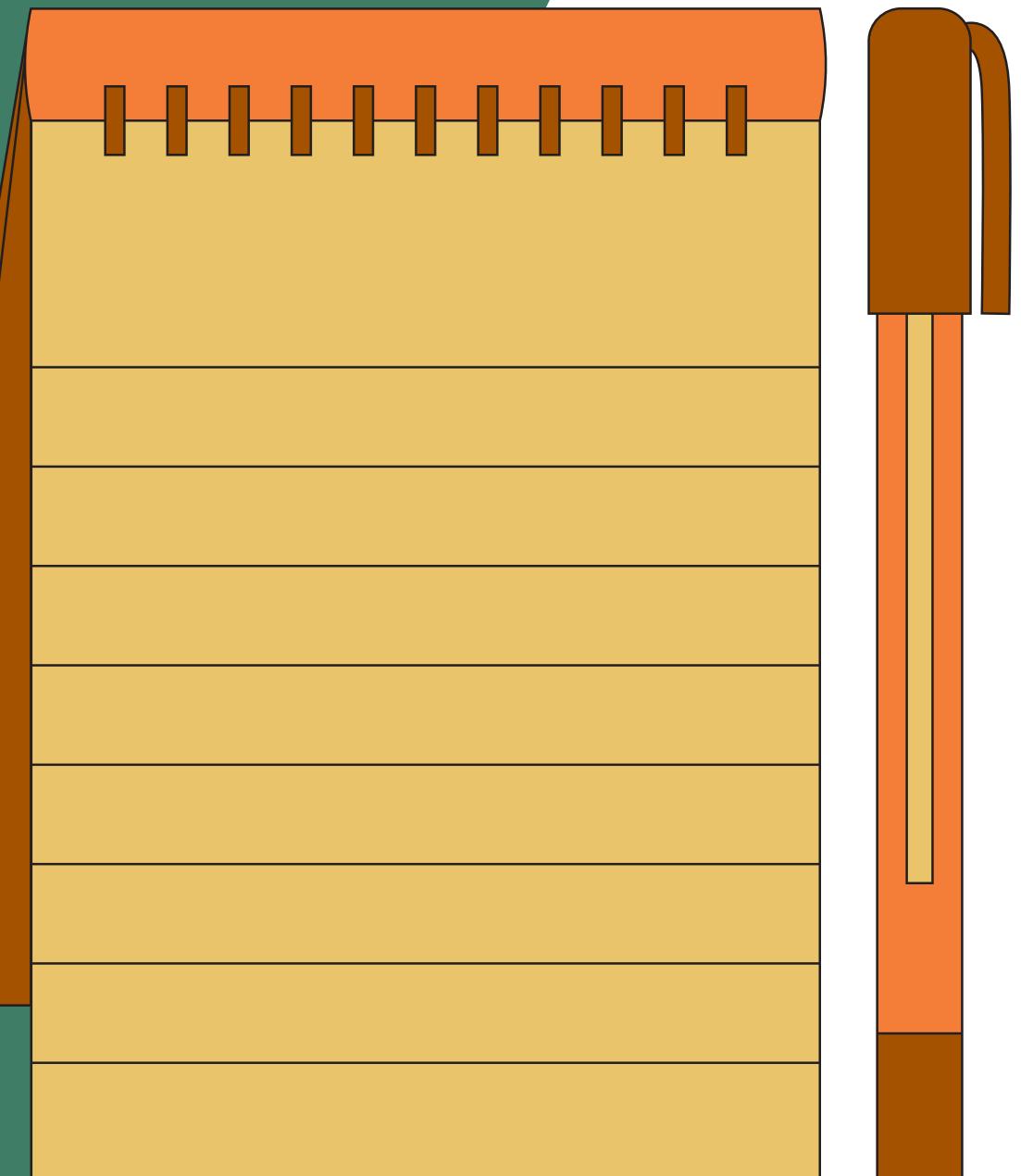
Bài báo này nghiên cứu hiệu suất của ChatGPT trong xác định thể loại tự động, một tác vụ phân loại văn bản khó cho con người. Nghiên cứu so sánh hiệu suất zero-shot của ChatGPT với bộ phân loại X-GENRE dựa trên mô hình XLM-RoBERTa. Kết quả cho thấy ChatGPT có hiệu suất ấn tượng, vượt trội hơn mô hình LLM tinh chỉnh trên bộ dữ liệu tiếng Anh. Đáng chú ý, mặc dù tiếng Slovenia là ngôn ngữ có nguồn lực thấp, hiệu suất của ChatGPT không kém hơn tiếng Anh nếu như việc hỏi đê xuất bằng tiếng Anh thay vì tiếng Slovenian.

**Điều này đặt ra câu hỏi liệu các chiến dịch gán nhãn lớn đã trở nên thừa và liệu có thể sử dụng ChatGPT để gán nhãn dữ liệu cho mục đích nghiên cứu hay không.**



Bài báo đặt ra câu hỏi liệu các "chiến dịch" chú thích lớn có trỏ nên dư thừa hay không, vì hiệu suất của ChatGPT cho thấy nó có thể được sử dụng để chú thích dữ liệu cho mục đích nghiên cứu. Cụ thể, trong phần 2, giới thiệu nhiệm vụ xác định thể loại tự động. Trình bày bộ dữ liệu được chú thích theo thể loại trong Phần 3, trong đó thử nghiệm mô hình ChatGPT và LLM tinh chỉnh, được trình bày trong Phần 4. Thảo luận về các dự đoán của ChatGPT trong Phần 5 và so sánh nó với mô hình tinh chỉnh trong Phần 6. Cuối cùng, trong Phần 7, kết thúc bài báo thảo luận về những phát hiện chính, đề xuất dự án tiếp theo.

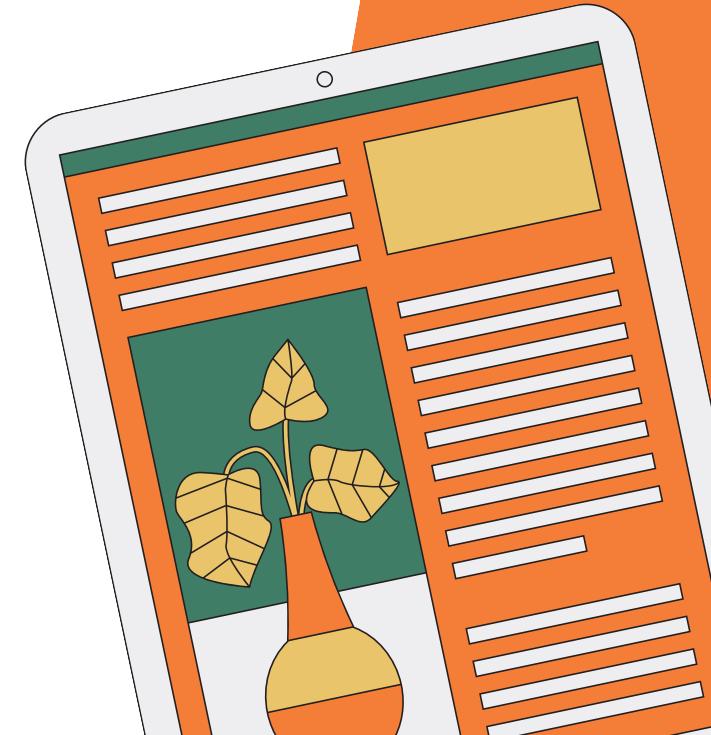
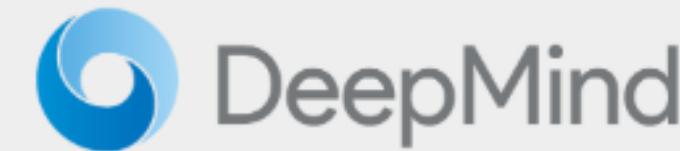
Trong phần này, có một số thứ ta cần lưu ý:



## GLUE (General Language Understanding Evaluation)

Benchmark là một bộ chuẩn đo lường được thiết kế để đánh giá và so sánh hiệu suất của các mô hình xử lý ngôn ngữ tự nhiên (NLP) trên một loạt các tác vụ hiểu ngôn ngữ chung. GLUE được giới thiệu vào năm 2018 bởi một nhóm các nhà nghiên cứu từ các tổ chức khác nhau như NYU, Facebook AI, và Google AI.

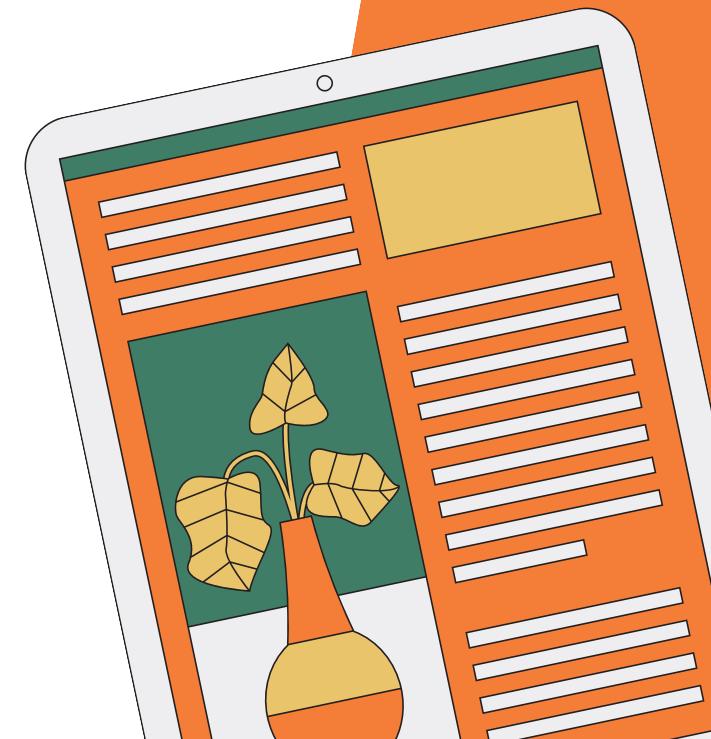
GLUE bao gồm 9 tác vụ NLP phổ biến, dựa trên các bộ dữ liệu khác nhau



## CoLA (Corpus of Linguistic Acceptability): Đánh giá mức độ ngữ pháp của các câu.

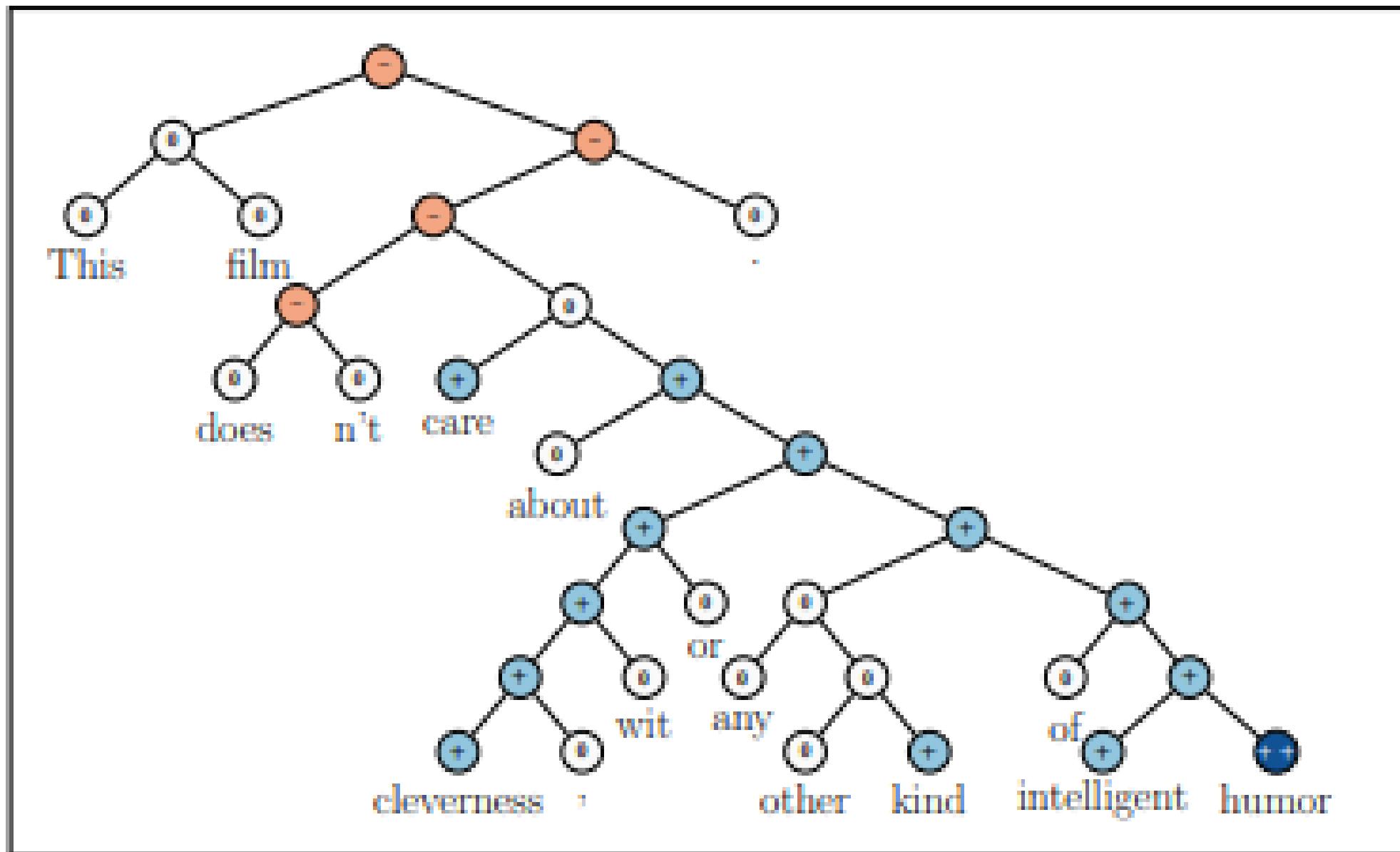
Là một bộ dữ liệu được tạo ra để đánh giá khả năng hiểu ngôn ngữ tự nhiên của các mô hình học máy. Bộ dữ liệu này bao gồm các câu tiếng Anh được thu thập từ nhiều nguồn, bao gồm các tài liệu học thuật về ngôn ngữ học và các ví dụ văn bản được tạo ra bởi các nhà ngôn ngữ học chuyên nghiệp.

Mục tiêu của CoLA là đánh giá khả năng phân biệt của các mô hình học máy giữa các câu ngôn ngữ học hợp lệ (câu có cấu trúc và ngữ nghĩa hợp lý) và các câu không hợp lệ (câu không có ý nghĩa hoặc vi phạm ngữ pháp tiếng Anh). Bằng cách sử dụng CoLA, các nhà nghiên cứu có thể đo lường hiệu suất của các mô hình học máy trong việc xử lý và hiểu ngôn ngữ tự nhiên, cũng như so sánh hiệu quả giữa các mô hình khác nhau.



**SST-2 (Stanford Sentiment Treebank v2):** Phân loại cảm xúc dựa trên đoạn văn bản. là một bộ dữ liệu được sử dụng để đánh giá các mô hình học máy trong việc phân tích cảm xúc. Bộ dữ liệu này được phát triển bởi nhóm nghiên cứu của Đại học Stanford và bao gồm các đánh giá phim được thu thập từ trang web Rotten Tomatoes.

SST-2 cung cấp nhãn cảm xúc cho từng câu trong các đánh giá phim, với mục đích phân loại chúng thành hai loại: tích cực hoặc tiêu cực. Các mô hình học máy được huấn luyện và kiểm tra trên SST-2 để đánh giá khả năng phân loại cảm xúc của chúng dựa trên văn bản đầu vào.



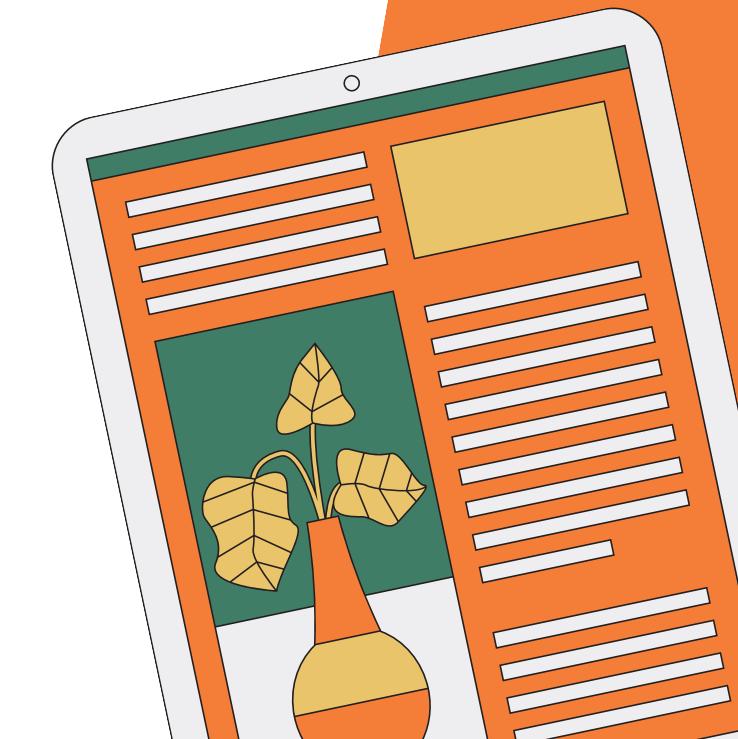


**MRPC (Microsoft Research Paraphrase Corpus):** Các mô hình học máy được huấn luyện trên MRPC để xác định liệu hai câu có phải là đồng nghĩa hay không.

Là một bộ dữ liệu được sử dụng trong các tác vụ xác định đồng nghĩa trong xử lý ngôn ngữ tự nhiên (NLP). Bộ dữ liệu này được phát triển bởi Microsoft Research và chứa các cặp câu được trích xuất từ tin tức trực tuyến. Mục đích của bộ dữ liệu là xác định liệu hai câu trong mỗi cặp có cùng nội dung hay không.

Đây là một tác vụ quan trọng trong NLP, vì nó giúp đánh giá khả năng của các mô hình trong việc hiểu và so sánh ngữ nghĩa của văn bản.

- **Sentence 1:** Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
- **Sentence 2:** Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- **Class:** 1 (true paraphrase)





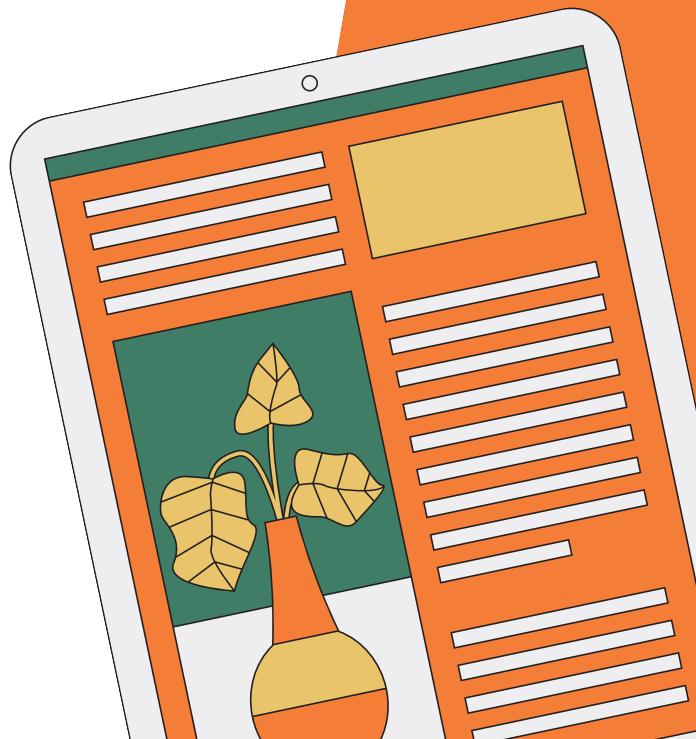
## **STS-B (Semantic Textual Similarity Benchmark): Tính điểm tương đồng ngữ nghĩa giữa hai câu.**

Là một bộ dữ liệu được sử dụng trong các tác vụ đánh giá độ tương tự ngữ nghĩa giữa các cặp câu trong xử lý ngôn ngữ tự nhiên (NLP). Bộ dữ liệu này bao gồm các cặp câu tiếng Anh được gán nhãn với điểm số thể hiện mức độ tương tự ngữ nghĩa của chúng, điểm số này nằm trong khoảng từ 0 (không tương tự) đến 5 (hoàn toàn tương tự).

Mục đích của STS-B là cung cấp một chuẩn mực để đánh giá khả năng của các mô hình học máy trong việc nắm bắt và so sánh ngữ nghĩa của các câu.

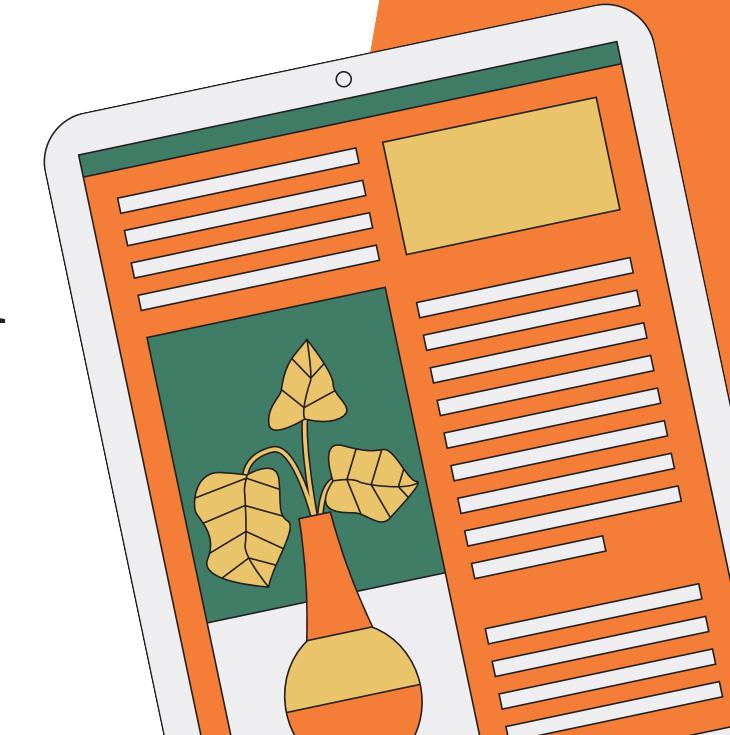


Score	English	Spanish
5/4	<p><i>The two sentences are completely equivalent, as they mean the same thing.</i></p> <p>The bird is bathing in the sink. Birdie is washing itself in the water basin.</p>	<p>El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.</p>
4	<p><i>The two sentences are mostly equivalent, but some unimportant details differ.</i></p> <p>In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.</p>	
3	<p><i>The two sentences are roughly equivalent, but some important information differs/missing.</i></p> <p>John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.</p>	<p>John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.</p>
2	<p><i>The two sentences are not equivalent, but share some details.</i></p> <p>They flew out of the nest in groups. They flew into the nest together.</p>	<p>Ellos volaron del nido en grupos. Volaron hacia el nido juntos.</p>
1	<p><i>The two sentences are not equivalent, but are on the same topic.</i></p> <p>The woman is playing the violin. The young lady enjoys listening to the guitar.</p>	<p>La mujer está tocando el violín. La joven disfruta escuchar la guitarra.</p>
0	<p><i>The two sentences are completely dissimilar.</i></p> <p>John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.</p>	<p>Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.</p>





**QQP (Quora Question Pairs):** Là một bộ dữ liệu trong xử lý ngôn ngữ tự nhiên (NLP), bao gồm các cặp câu hỏi được trích xuất từ cộng đồng trực tuyến Quora. Mục tiêu của bộ dữ liệu này là xác định liệu hai câu hỏi có cùng nội dung hay không, tức là liệu chúng có đang hỏi về cùng một chủ đề hay không.



Bộ dữ liệu QQP bao gồm hàng trăm nghìn cặp câu hỏi tiếng Anh, mỗi cặp đều được gán nhãn là "trùng lặp" (nếu cả hai câu hỏi cùng nội dung) hoặc "không trùng lặp" (nếu cả hai câu hỏi không cùng nội dung). QQP thường được sử dụng để đánh giá khả năng của các mô hình học máy trong việc nắm bắt ngữ nghĩa của các câu hỏi và xác định sự tương tự giữa chúng, nhằm giúp giảm bớt việc trùng lặp câu hỏi trên các diễn đàn trực tuyến và cải thiện chất lượng của các câu trả lời.



	<b>id</b>	<b>qid1</b>	<b>qid2</b>	<b>question1</b>	<b>question2</b>	<b>is_duplicate</b>
0	0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0
2	2	5	6	How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
3	3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when $23^{24}$ is divided by 24,23?	0
4	4	9	10	Which one dissolve in water quickly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1



<b>id</b>	<b>qid1</b>	<b>qid2</b>	<b>question1</b>	<b>question2</b>	<b>is_duplicate</b>
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

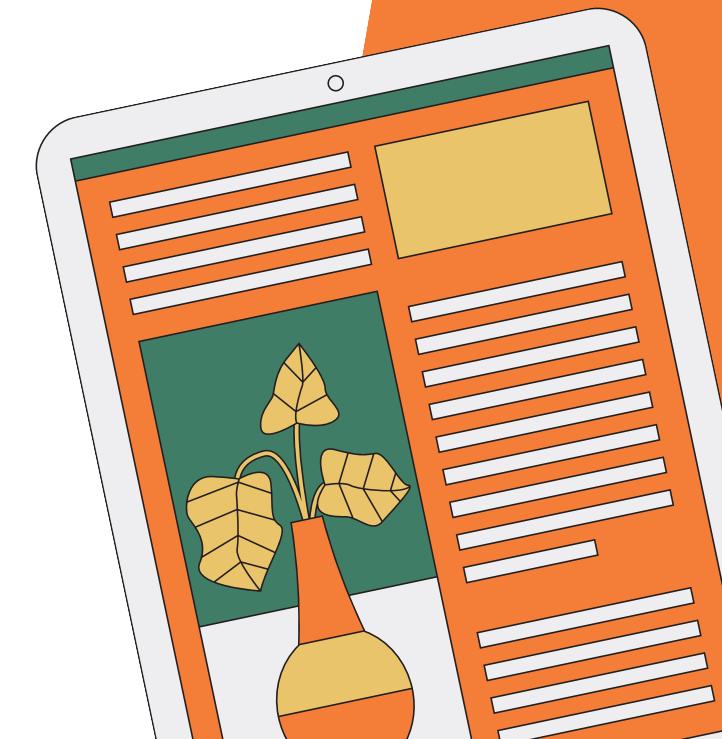
# MNLI (Multi-Genre Natural Language Inference): Xác định mối quan hệ logic giữa hai câu (đồng thuận, đối lập, trung lập).



Là một bộ dữ liệu lớn dành cho xử lý ngôn ngữ tự nhiên (NLP) trong nhiệm vụ suy luận ngôn ngữ tự nhiên (NLI). NLI là quá trình xác định mối quan hệ logic giữa hai câu văn: một câu tiên đề và một câu giả thuyết.

MNLI bao gồm các cặp câu được thu thập từ nhiều nguồn văn bản thuộc nhiều thể loại khác nhau, như văn học, tin tức, hội thoại, và văn bản học thuật. Mỗi cặp câu đều được gán một trong ba nhãn sau:

- Đồng thuận (entailment):** Nếu câu giả thuyết đúng, thì câu tiên đề cũng đúng.
- Đối lập (contradiction):** Nếu câu giả thuyết đúng, thì câu tiên đề sai.
- Trung lập (neutral):** Mối quan hệ giữa câu giả thuyết và câu tiên đề không rõ ràng, hoặc không có liên quan.



Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

Premise	Hypothesis	Label
A black race car starts up in front of a crowd of people.	A man is driving a lonely road.	Contradiction
A soccer game with multiple males playing.	Some men are playing a sport.	Entailment
A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.	Neutral

A man inspects the uniform of a figure in some East Asian country.

**contradiction**

C C C C C

The man is sleeping

An older and younger man smiling.

**neutral**

N N E N N

Two men are smiling and laughing at the cats playing on the floor.

A black race car starts up in front of a crowd of people.

**contradiction**

C C C C C

A man is driving down a lonely road.

A soccer game with multiple males playing.

**entailment**

E E E E E

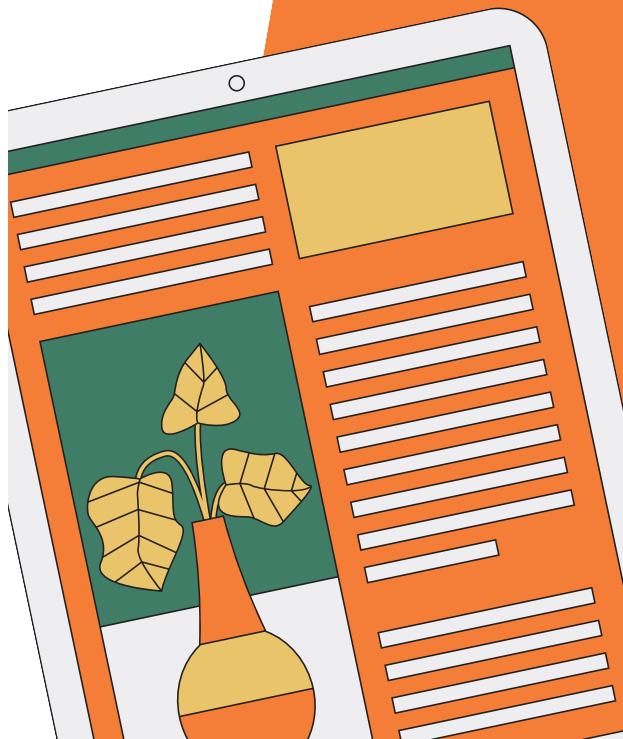
Some men are playing a sport.

A smiling costumed woman is holding an umbrella.

**neutral**

N N E C N

A happy woman in a fairy costume holds an umbrella.



# QNLI (Question-answering Natural Language Inference): Tác vụ hỏi đáp dựa trên tập dữ liệu SQuAD.



Là một bộ dữ liệu được tạo ra để đánh giá các mô hình xử lý ngôn ngữ tự nhiên (NLP) trong nhiệm vụ suy luận ngôn ngữ tự nhiên (NLI) với bối cảnh câu hỏi và trả lời. QNLI được chuyển đổi từ bộ dữ liệu SQuAD (Stanford Question Answering Dataset), một bộ dữ liệu phổ biến trong nhiệm vụ trả lời câu hỏi dựa trên đoạn văn.

Trong QNLI, mục tiêu là xác định liệu câu trả lời có xuất phát từ câu tiền đề (đoạn văn) hay không, khi cho trước một câu hỏi và một câu trả lời để xuất. Mỗi cặp câu hỏi và câu trả lời được gán nhãn "entailment" nếu câu trả lời đúng theo câu tiền đề, và "not\_entailment/contradiction" nếu câu trả lời không đúng theo câu tiền đề.

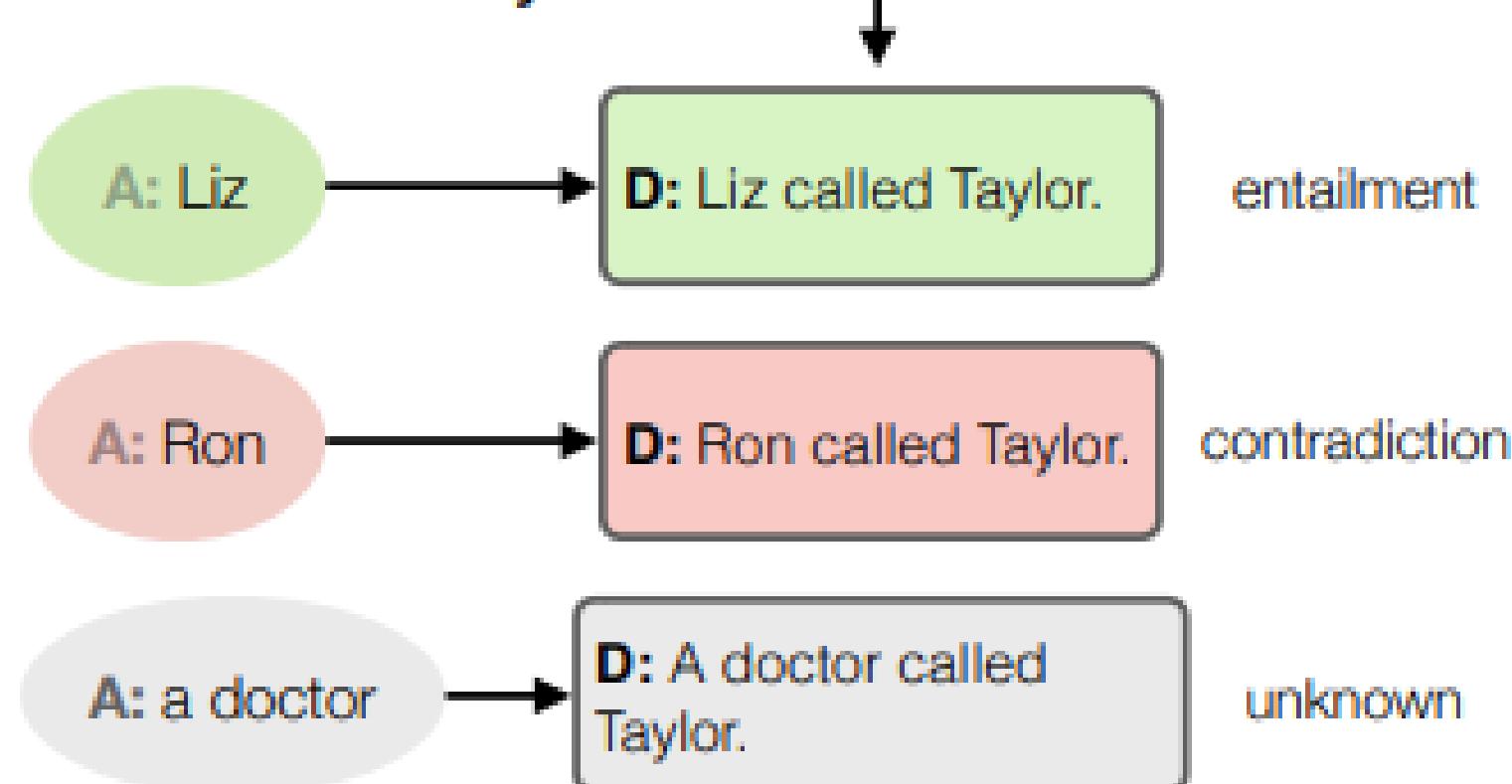


Bằng cách sử dụng bộ dữ liệu QNLI, các nhà nghiên cứu có thể đánh giá khả năng của các mô hình NLP trong việc suy luận dựa trên câu hỏi và câu trả lời, và cải thiện hiệu suất của các mô hình trong nhiệm vụ trả lời câu hỏi.

P: Taylor is a journalist [...]. She was playing golf with Ron when her phone rang. It was Liz, her mother's friend. [...]

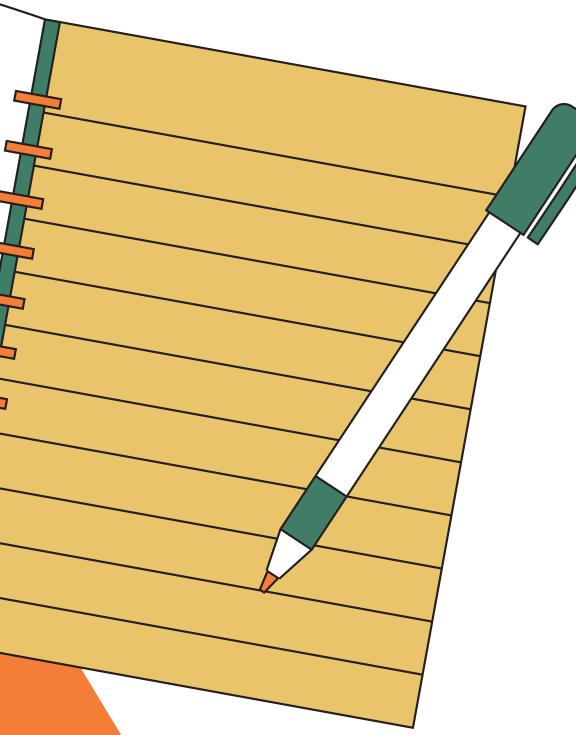
Q: Who called Taylor?

QA2D



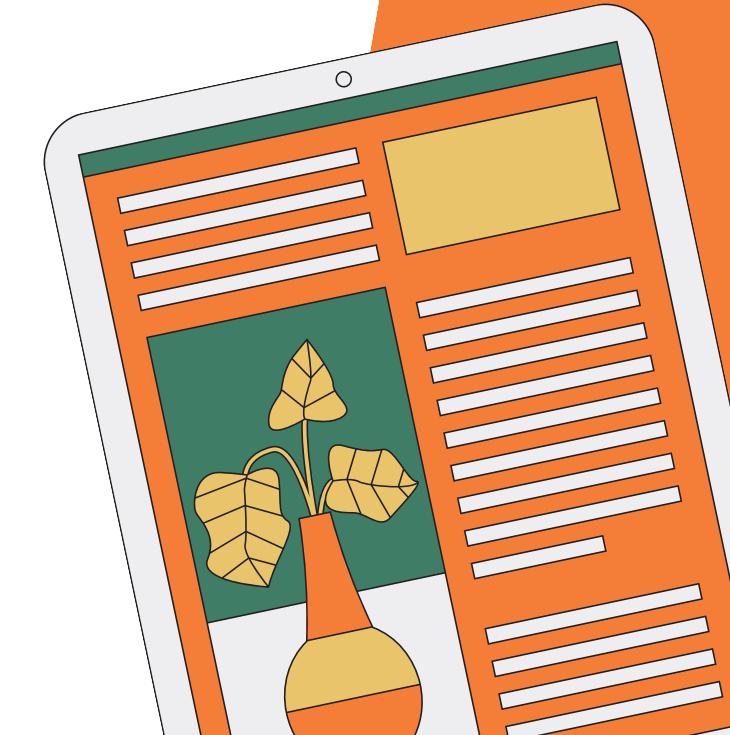
The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?  
Answer: through contact with Persian traders



**RTE (Recognizing Textual Entailment):** Xác định liệu một câu có kết luận từ câu kia hay không. Là một nhiệm vụ trong xử lý ngôn ngữ tự nhiên (NLP) mà mục tiêu là xác định liệu một đoạn văn (giả thuyết) có thể được suy ra từ một đoạn văn khác (tiền đề) hay không. Nhiệm vụ này đánh giá khả năng của các mô hình NLP trong việc hiểu và suy luận ngữ nghĩa từ văn bản.

Trong RTE, mỗi cặp đoạn văn được gán nhãn "entailment" nếu giả thuyết có thể được suy ra từ tiền đề, và "not\_entailment" nếu giả thuyết không thể được suy ra từ tiền đề.





### Ví dụ 1:

Tiên đề: "Công ty XYZ thông báo tăng trưởng doanh thu quý này."

Giả thuyết: "Doanh thu của công ty XYZ đã tăng trong quý này."

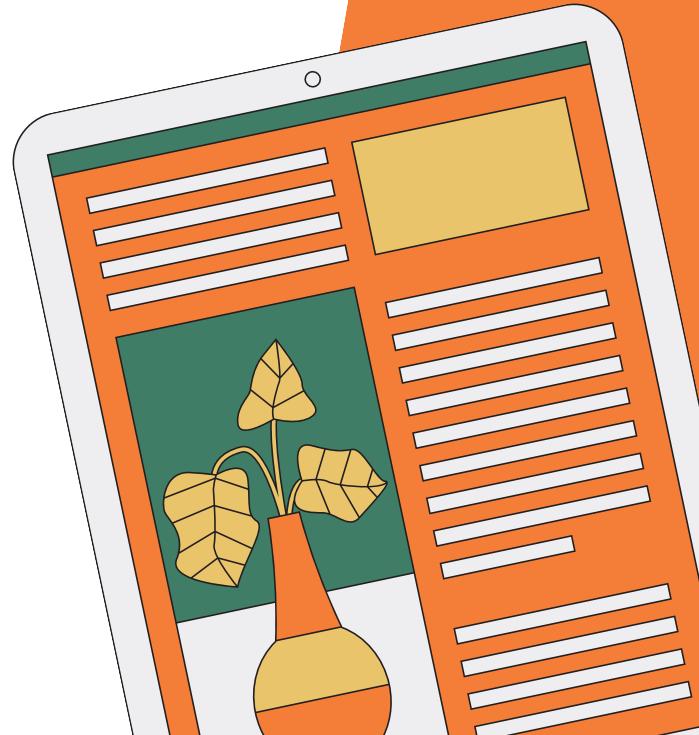
Nhãn: Entailment

### Ví dụ 2:

Tiên đề: "Hàng loạt các cửa hàng VinMart đóng cửa"

Giả thuyết: "Messi người hùng Chile đã thâu hết chuỗi Điện máy xanh"

Nhãn: Not\_Entailment or Unknown

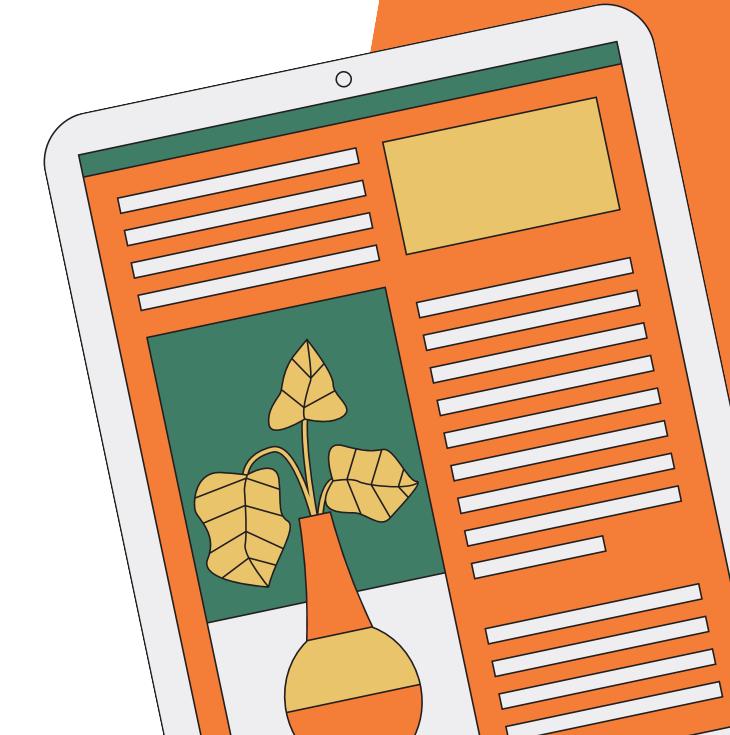


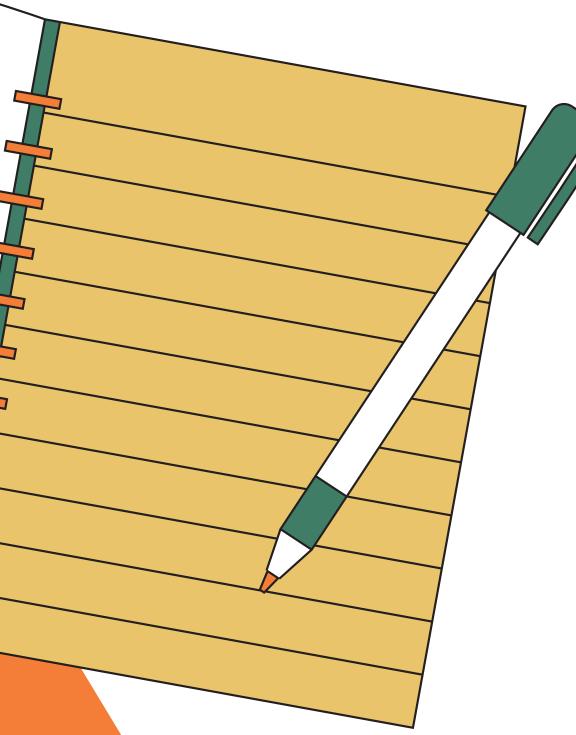


## WNLI (Winograd Natural Language Inference): Đọc hiểu dựa trên các câu Winograd

**Schema.**Bài toán Winograd Schema được thiết kế để kiểm tra khả năng suy luận ngữ nghĩa và hiểu biết ngôn ngữ tự nhiên của các mô hình NLP bằng cách đưa ra các câu có đại từ nhân xưng mơ hồ và yêu cầu mô hình xác định thực thể nào đại từ nhân xưng đó đang chỉ đến.

Trong WNLI, mỗi ví dụ bao gồm một cặp câu (tiền đề và giả thuyết) cùng với một nhãn chỉ ra liệu giả thuyết có được suy ra từ tiền đề hay không. Mục tiêu của nhiệm vụ là xác định liệu giả thuyết có thể được suy ra từ tiền đề dựa trên ngữ nghĩa của các câu.





### Ví dụ 1:

Tiên đề: "Cô bé đặt chiếc bánh vào chiếc hộp vì nó quá nhỏ."

Giả thuyết: "Cô bé đặt chiếc bánh vào chiếc hộp vì chiếc bánh quá nhỏ."

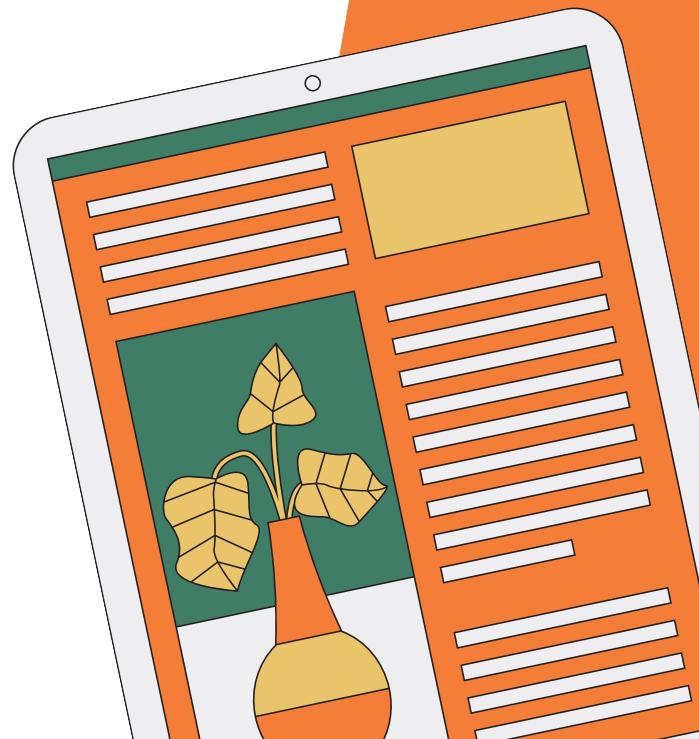
Nhãn: Entailment

### Ví dụ 2:

Tiên đề: "Zamorano: Nếu Messi ở Chile, anh ấy có thể được tạc tượng"

Giả thuyết: "Messi sút hỏng quả phạt đền 11m CK Copa America 2016."

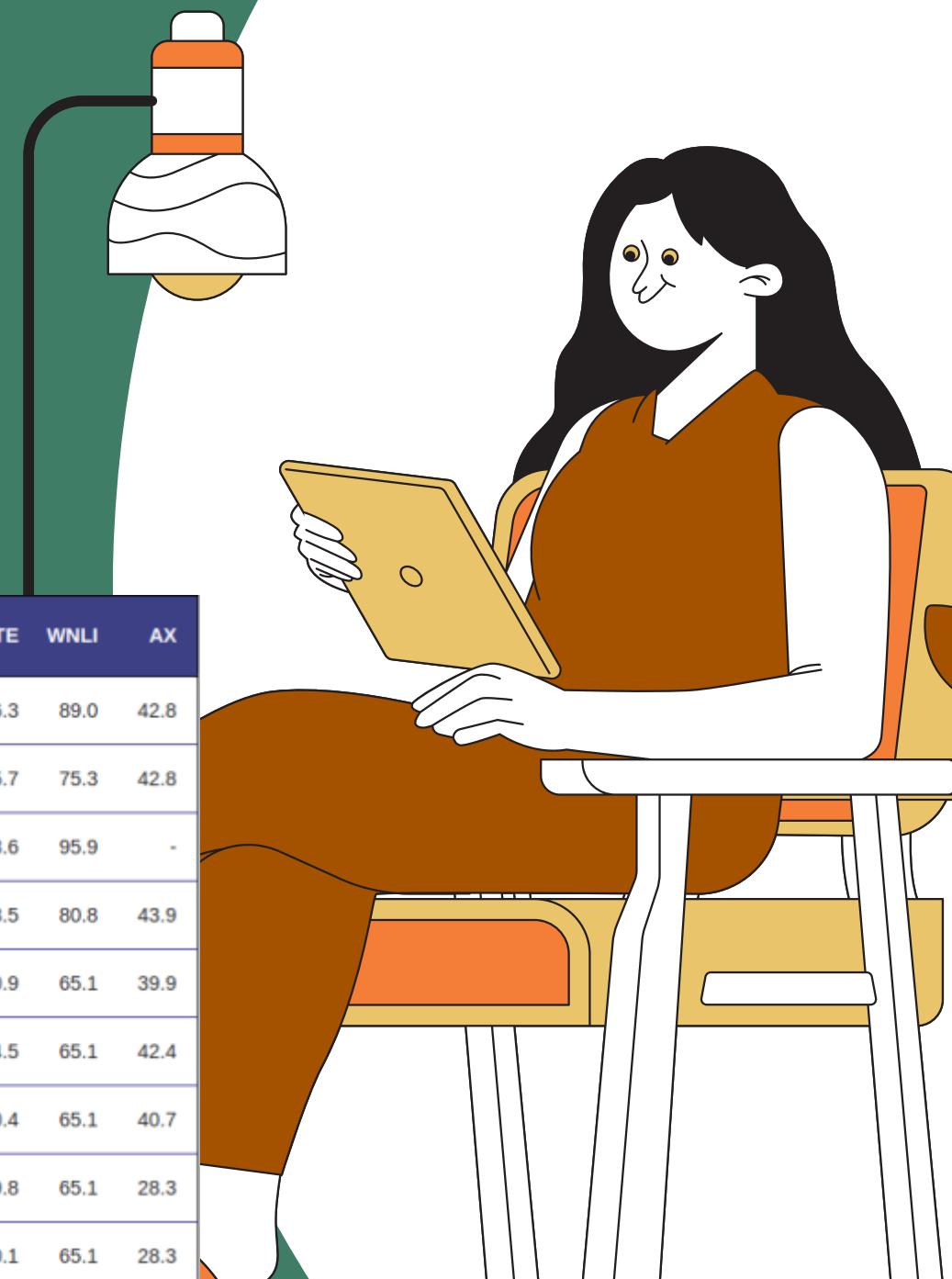
Nhãn: Entailment



Để đánh giá hiệu suất, các mô hình được đưa vào GLUE leaderboard, nơi chúng được xếp hạng dựa trên điểm số trung bình trên tất cả các tác vụ. Điểm số này được gọi là GLUE Score.

Rank Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP
1	T5 Team - Google	<a href="#">T5</a>	89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4
2	ALBERT-Team Google Language	<a href="#">ALBERT (Ensemble)</a>	89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5
+ 3	王玮	<a href="#">ALICE v2 large ensemble (Alibaba DAMO NLP)</a>	89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7

Rank Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
4	Microsoft D365 AI & UMD	<a href="#">FreeLB-RoBERTa (ensemble)</a>	87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
5	Facebook AI	<a href="#">RoBERTa</a>	85.1	65.4	95.6	91.2/88.1	89.6/89.0	73.7/89.9	87.9	87.4	96.0	85.7	75.3	42.8
6	XLNet Team	<a href="#">XLNet-Large (ensemble)</a>	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
+ 7	Microsoft D365 AI & MSR AI	<a href="#">MT-DNN-ensemble</a>	86.3	68.6	95.2	92.6/90.2	91.1/90.6	74.4/90.7	88.2	87.9	95.7	83.5	80.8	43.9
8	GLUE Human Baselines	<a href="#">GLUE Human Baselines</a>	83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
9	Stanford Hazy Research	<a href="#">Snorkel MeTaL</a>	82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
10	XLM Systems	<a href="#">XLM (English only)</a>	82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
			82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
			82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
			80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
			80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
			79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7





GLUE Benchmark đã giúp thúc đẩy sự phát triển và cạnh tranh trong lĩnh vực NLP, khiến các nhà nghiên cứu không ngừng tìm kiếm các mô hình và kỹ thuật mới để cải thiện hiệu suất. Tuy nhiên, do sự tiến bộ nhanh chóng trong NLP, GLUE đã được thay thế bởi SuperGLUE, một bộ chuẩn đo lường mới và khó khăn hơn, để tiếp tục thúc đẩy sự phát triển trong lĩnh vực này.



Ở đây, tác giả đã lấy GLUE benchmark so sánh ChatGPT với các mô hình ngôn ngữ tinh chỉnh BERT và RoBERTa



facebook Artificial Intelligence

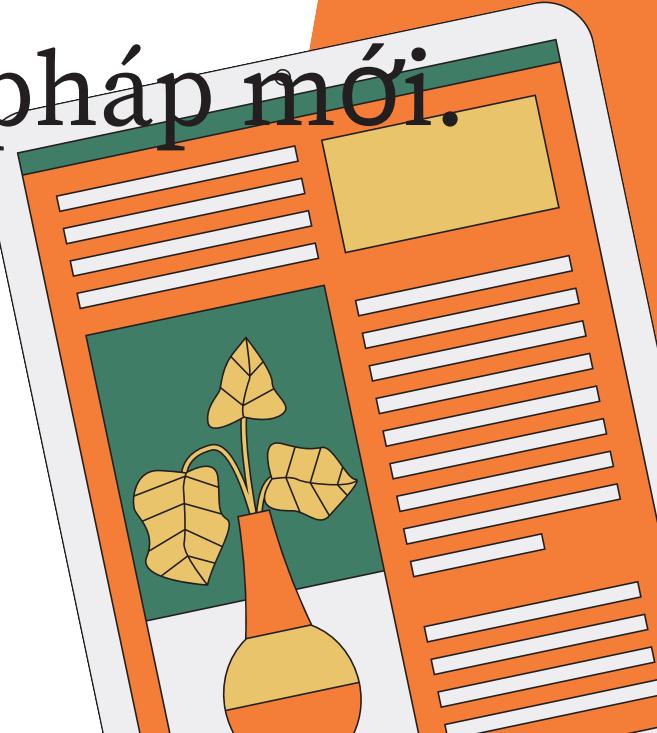


DeepMind

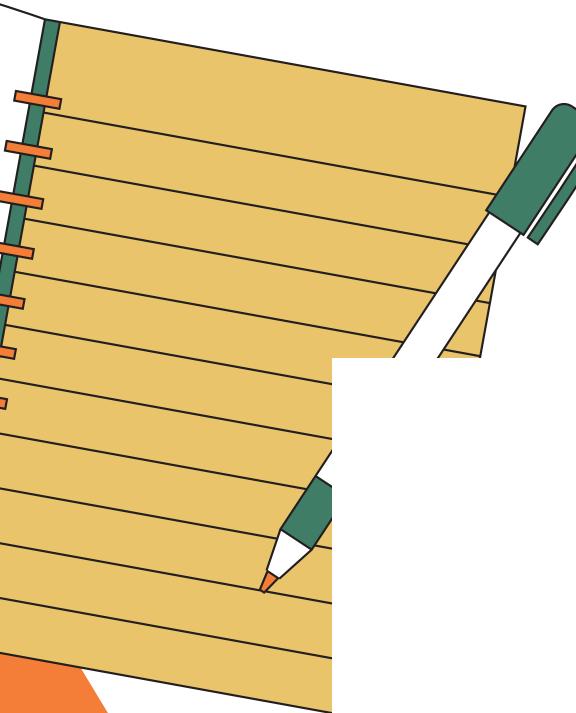


Thứ hai, thuật ngữ "SOTA" là viết tắt của "State Of The Art", được sử dụng để chỉ các mô hình hoặc phương pháp hàng đầu trong một lĩnh vực nghiên cứu hoặc công nghệ cụ thể. Trong ngữ cảnh của học máy và xử lý ngôn ngữ tự nhiên (NLP), các **mô hình SOTA** là những mô hình đạt được hiệu suất cao nhất hoặc xuất sắc nhất trên một tác vụ hoặc bộ chuẩn đo lường cụ thể.

Khi nói đến các **mô hình SOTA**, chúng ta thường nghĩ đến các mô hình đã được huấn luyện và tinh chỉnh để đạt được kết quả tốt nhất trong một lĩnh vực nghiên cứu hoặc tác vụ cụ thể. Các mô hình này thường được công nhận rộng rãi trong cộng đồng nghiên cứu và phát triển công nghệ, đồng thời được sử dụng làm điểm chuẩn để so sánh với các mô hình và phương pháp mới.



# Thứ ba, WMT22 và GINCO datasets là hai bộ dữ liệu được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) cho các tác vụ khác nhau.



Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1584–1594  
Marseille, 20–25 June 2022  
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

## MUNI-NLP Systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian Machine Translation @ WMT22

Edoardo Signoroni and Pavel Rychly

Faculty of Informatics  
Masaryk University

e.signoroni@mail.muni.cz, pary@fi.muni.cz

### Abstract

We describe our neural machine translation systems for the WMT22 shared task on unsupervised MT and very low resource supervised MT. We submit supervised NMT systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian translation in both directions. By using a novel tokenization algorithm, data augmentation techniques, such as Data Diversification (DD), and parameter optimization we improve on our baselines by 10.5-10.77 BLEU for Lower Sorbian-German and by 1.52-1.88 BLEU for Lower Sorbian-Upper Sorbian.

### Introduction

This paper describes our Machine Translation (MT) systems for the WMT22 shared task on "Unsupervised MT and Very Low Resource Supervised MT"<sup>1</sup>, which features translation between Lower Sorbian, Upper Sorbian, and German. Lower (*dsb*) and Upper Sorbian (*hsb*) are Slavic minority languages spoken in the Eastern part of Germany with 7,000 and 30,000 native speakers respectively. Text data for these languages collected and made available by the Sorbian Institute and the Witaj Language Centre (Libovický and Fraser, 2021).

We submit systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian in both translation directions. We focused on the supervised ap-

We employed HFT since it aims to obtain more meaningful subword dictionaries, while DD was chosen because it does not involve additional data apart from the original parallel corpus. Both these techniques are relevant when working with a limited amount of data.

This paper is structured as follows: Section 1 summarizes the data used in training; Section 2 outlines our methodology, introducing our novel tokenizer and the models we used; Section 3 sums up our final systems, while Section 4 relates and discusses the results of our experiments; Section 5 contains some final remarks.

### 1 Data

We experiment with Lower Sorbian-German and Lower Sorbian-Upper Sorbian translation, using only the parallel data provided for each pair.

The parallel data for the *dsb-de* and the *dsb-hsb* pairs consist of ~40k and ~62k sentences respectively. We use only these data, as the approach we decided to follow does not need additional data. After applying this method, our final corpus size for training is ~360k sentences for *dsb-de*, and ~560k for *dsb-hsb*.

### 2 Methodology

In this section, we first present briefly our novel to-

## The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild

Taja Kuzman, Peter Rupnik, Nikola Ljubešić

Jožef Stefan Institute  
Jamova cesta 39, 1000 Ljubljana  
taja.kuzman@ijs.si, peter.rupnik@ijs.si, nikola.ljubesic@ijs.si

### Abstract

This paper presents a new training dataset for automatic genre identification GINCO, which is based on 1,125 crawled Slovenian web documents that consist of 650,000 words. Each document was manually annotated for genre with a new annotation schema that builds upon existing schemata, having primarily clarity of labels and inter-annotator agreement in mind. The dataset consists of various challenges related to web-based data, such as machine translated content, encoding errors, multiple contents presented in one document etc., enabling evaluation of classifiers in realistic conditions. The initial machine learning experiments on the dataset show that (1) pre-Transformer models are drastically less able to model the phenomena, with macro F1 metrics ranging around 0.22, while Transformer-based models achieve scores of around 0.58, and (2) multilingual Transformer models work as well on the task as the monolingual models that were previously proven to be superior to multilingual models on standard NLP tasks.

**Keywords:** automatic genre identification, web genres, genre classification schema, web corpora, Slovenian language

### 1. Introduction

With the arrival of the Web, it has become significantly easier to collect very large corpora that fuel innovation and creation of advanced resources and language technologies. However, contrary to the traditionally collected corpora, web corpora are built in an automated way which limits the control over the contents that constitute the final corpus (Baroni et al., 2009). One of the post hoc evaluation methods to investigate the corpus composition and quality, and to enrich the corpus with important metadata is automatic genre identification (AGI). This method focuses on genres as text categories based on the author's purpose, the socially recognized function of a document and/or the conventional patterns of form, following the definition by Orlíkowska and Yates (1994). As this research is a part of the MaCoCu<sup>1</sup> project that aims to collect large corpora for under-resourced languages, our main purpose is to provide a classifier that would efficiently identify genres in web corpora for Slovenian and other languages, which would allow an in-depth analysis of the quality and composition of the newly provided corpus.

### 2. Related work

Despite the considerable benefits of automatic genre identification (AGI), no established classification exists (Sharoff, 2010). The genre researchers are not consistent in the use of terminologies and their refer to genera-

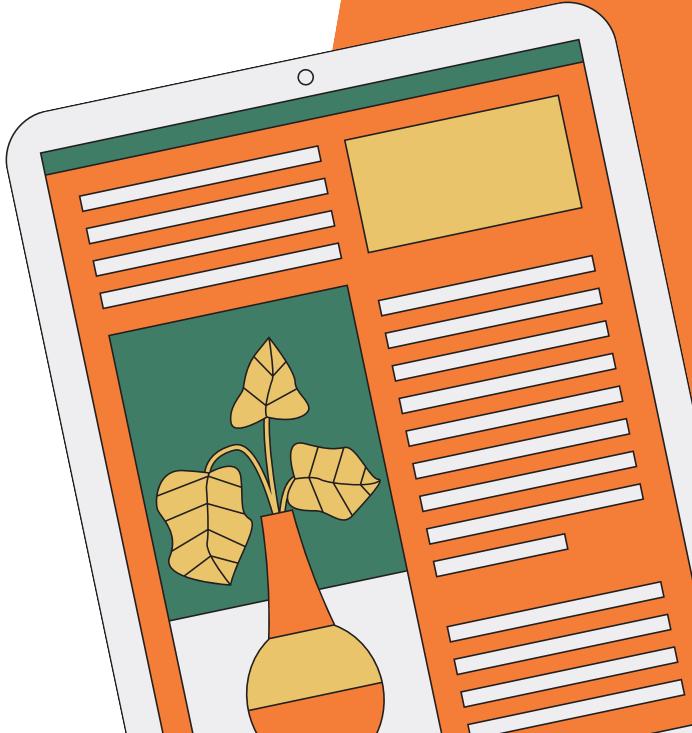


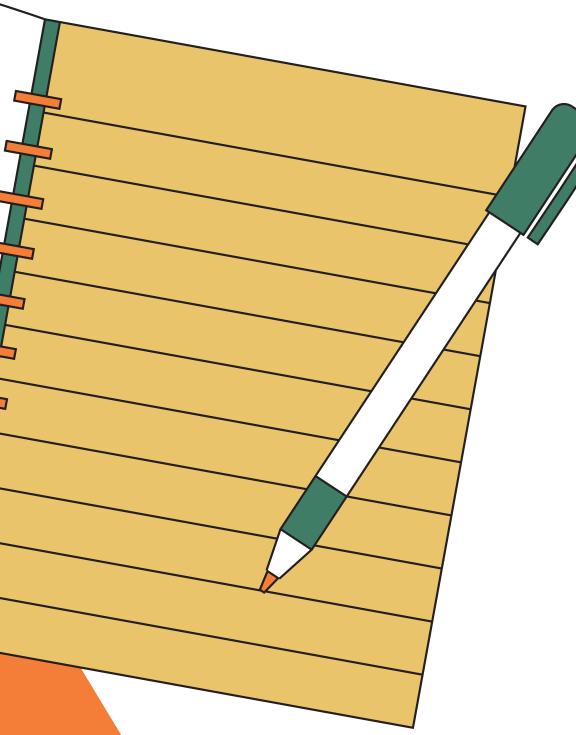
**WMT22:** WMT là viết tắt của "Workshop on Machine Translation", một hội thảo hàng năm tập trung vào các vấn đề liên quan đến dịch máy. WMT22 đề cập đến kỳ hội thảo năm 2022. Trong khuôn khổ của hội thảo này, các nhà nghiên cứu tham gia các cuộc thi dịch máy để đánh giá và so sánh hiệu suất của các mô hình dịch máy khác nhau. WMT22 dataset là bộ dữ liệu được sử dụng trong cuộc thi dịch máy của hội thảo WMT năm 2022, bao gồm các cặp câu nguồn và mục tiêu trong nhiều ngôn ngữ khác nhau.

**GINCO dataset:** GINCO (Genre Identification and Clustering of Corpora) là một bộ dữ liệu được sử dụng để nghiên cứu và đánh giá hiệu suất của các mô hình trong tác vụ xác định thể loại văn bản tự động. GINCO dataset bao gồm các văn bản trong nhiều ngôn ngữ khác nhau, được gán nhãn với các thể loại văn bản tương ứng. Bộ dữ liệu này hỗ trợ việc huấn luyện và kiểm tra các mô hình phân loại văn bản đa ngôn ngữ để xác định thể loại của các văn bản mới.



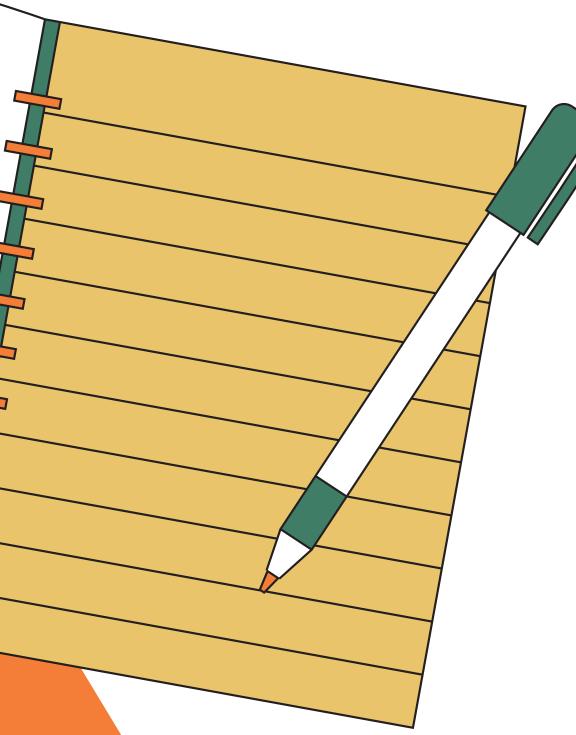
Cả hai bộ dữ liệu này đều quan trọng và hữu ích trong việc nâng cao hiệu suất của các mô hình NLP và giúp cải thiện các ứng dụng thực tế của công nghệ này.





### **3. Nhận dạng thẻ loại tự động:**



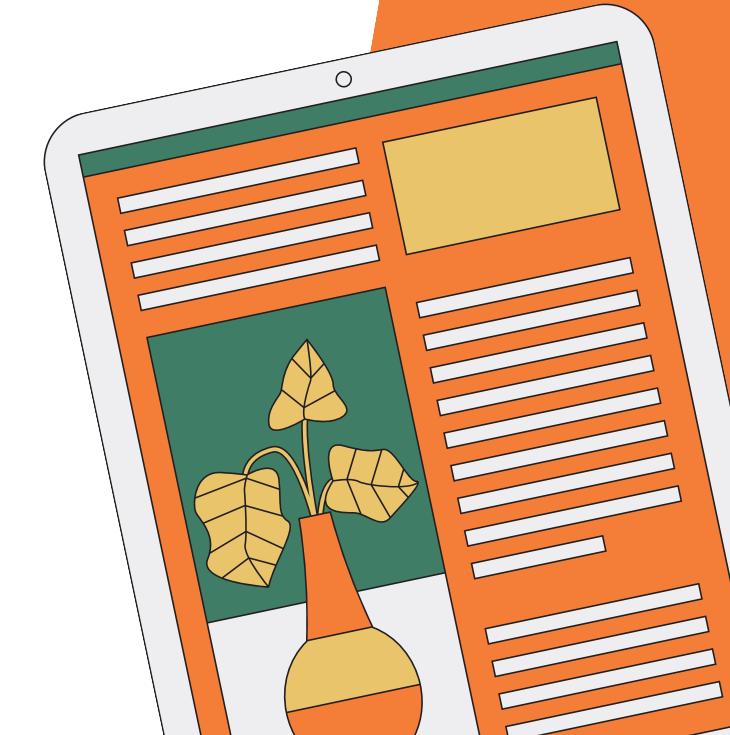


# Automatic Genre Identification (AGI) là gì?





Đó là một tác vụ trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) liên quan đến việc tự động xác định thể loại của một văn bản. Thể loại văn bản là một phân loại dựa trên nội dung, chức năng, cấu trúc và/hoặc phong cách của văn bản. Ví dụ, một số thể loại phổ biến bao gồm tin tức, tiểu luận, thơ, tiểu thuyết, hướng dẫn và báo cáo khoa học.



Tác vụ AGI đòi hỏi phát triển các mô hình và thuật toán có khả năng phân tích văn bản và xác định thể loại của chúng một cách chính xác. Điều này đặc biệt quan trọng trong các ứng dụng như lọc thông tin, gợi ý nội dung, tổ chức và phân loại văn bản lớn.

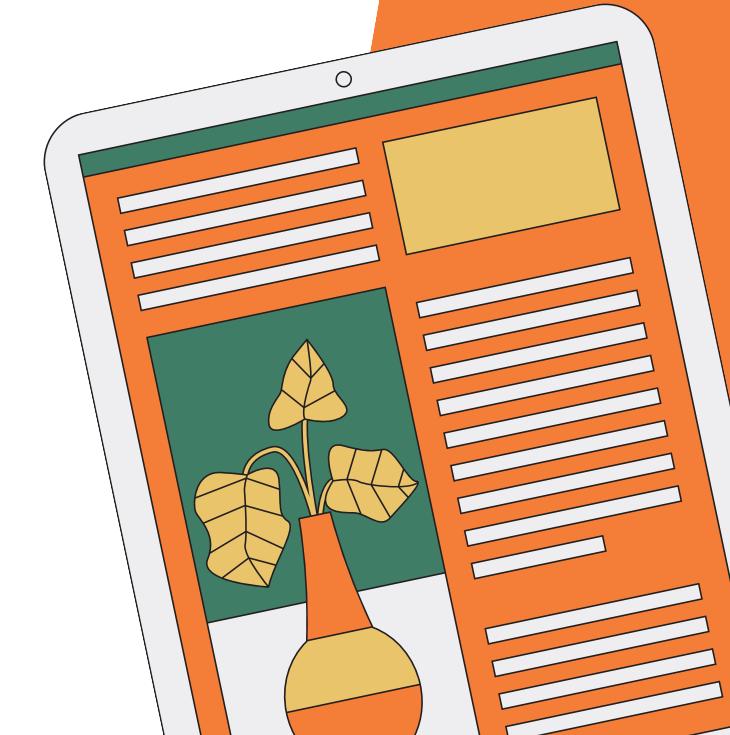
Các kỹ thuật phổ biến để giải quyết tác vụ AGI bao gồm các phương pháp học máy truyền thống như Naive Bayes, học máy vector hỗ trợ (SVM) và cây quyết định, cũng như các mô hình Transformer như BERT, GPT và RoBERTa.

Để huấn luyện và đánh giá hiệu suất của các mô hình trong tác vụ AGI, các nhà nghiên cứu sử dụng các bộ dữ liệu đã được gán nhãn thể loại văn bản, chẳng hạn như bộ dữ liệu GINCO.





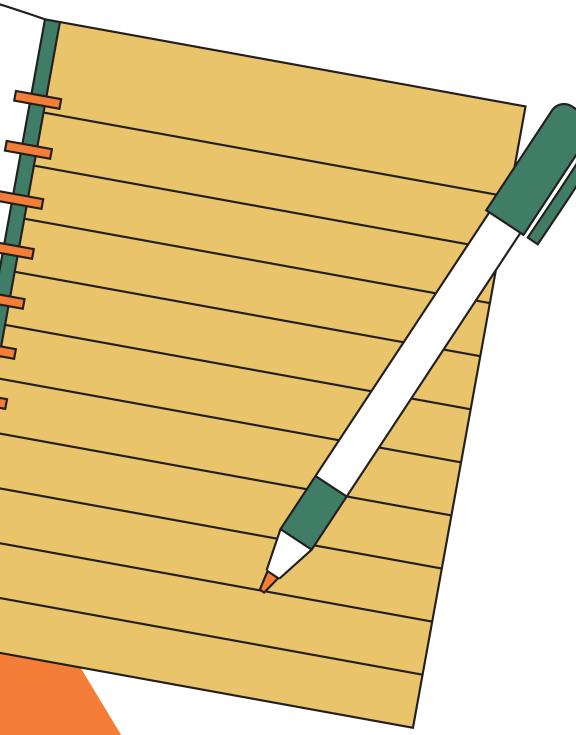
Khác biệt so với phát hiện chủ đề, thể loại không thể phân loại chỉ dựa trên thông tin từ vựng. Các mô hình cần tìm ra mẫu mã cao hơn trong văn bản, thường dựa trên đặc trưng văn bản hoặc cú pháp không liên quan đến chủ đề. Phân loại thể loại văn bản là một tác vụ khó, vì văn bản có thể là ví dụ đặc trưng hơn hoặc ít đặc trưng hơn cho thể loại, hiển thị dấu hiệu của nhiều lớp hoặc thiếu dấu hiệu của bất kỳ thể loại nào.



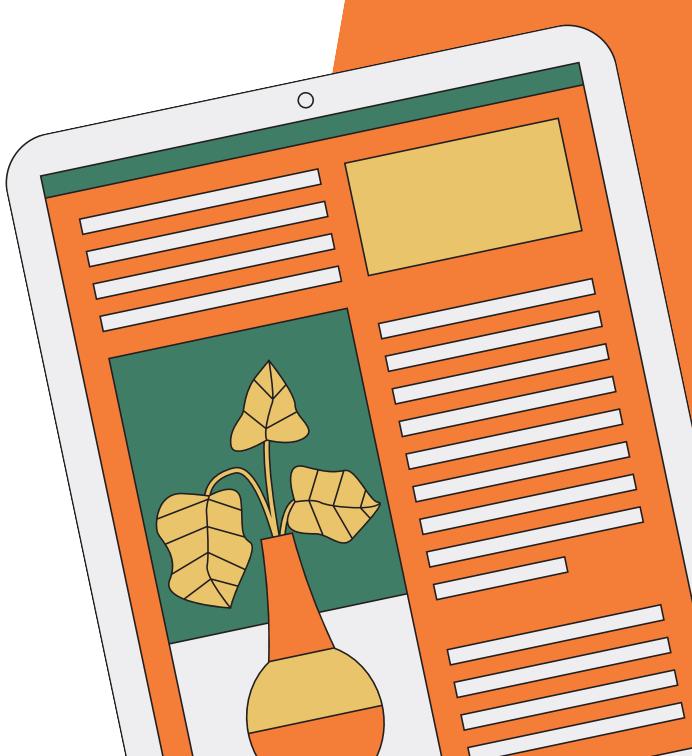
Các phương pháp phi mạng nơron đã chứng minh là quá phụ thuộc vào tập dữ liệu và không thể tổng quát hóa cho các tập dữ liệu chưa từng thấy. Tuy nhiên, sự xuất hiện của các mô hình Transformer dựa trên mạng nơron sâu đã mang lại bước đột phá. Các mô hình giống BERT được tiền huấn luyện trên lượng lớn văn bản và tinh chỉnh cho tác vụ xác định thể loại, cho thấy chúng có thể xác định thể loại trên các tập dữ liệu và ngôn ngữ chưa từng thấy, thậm chí chỉ cần huấn luyện trên 1.000 văn bản. Tuy nhiên, các mô hình vẫn cần văn bản được gán nhãn thủ công, một công việc tốn thời gian và chi phí.

<b>Labels</b>	<b>EN-GINCO</b>	<b>GINCO</b>	<b>X-GENRE training</b>
Information/Explanation	25%	24%	17%
Promotion	22%	17%	16%
Opinion/Argumentation	18%	11%	14%
News	18%	29%	19%
Other	6%	7%	4%
Forum	6%	5%	8%
Instruction	5%	5%	12%
Legal	0%	1%	4%
Prose/Lyrical	0%	1%	6%

**Bảng 1: Phân phôi nhãn trong các tập dữ liệu kiểm tra tiếng Anh (EN-GINCO), tập dữ liệu kiểm tra tiếng Slovenia (GINCO) và tập dữ liệu được sử dụng để huấn luyện bộ phân loại X-GENRE. Những con số này giúp hiểu rõ hơn về cấu trúc và đặc điểm của từng tập dữ liệu và có thể được sử dụng để đánh giá hiệu suất của mô hình trên các tác vụ phân loại thể loại.**



## 4. Tập dữ liệu được gán nhãn thể loại:





**Genre-annotated datasets (tập dữ liệu được gán nhãn thể loại)** là những tập dữ liệu văn bản mà mỗi văn bản đều được gán nhãn với một hoặc nhiều thể loại văn bản. Thể loại văn bản thường được xác định dựa trên mục đích của tác giả, chức năng và hình thức thông thường của văn bản, chẳng hạn như tin tức, pháp lý, quảng cáo, ý kiến, hướng dẫn, văn chương, và nhiều hơn nữa.



Các tập dữ liệu này thường được sử dụng trong các tác vụ liên quan đến phân loại thể loại tự động (Automatic Genre Identification) để huấn luyện, tinh chỉnh và đánh giá các mô hình phân loại văn bản theo thể loại. Việc gán nhãn thể loại giúp các mô hình học cách nhận diện các đặc trưng và mẫu mã riêng biệt của từng thể loại văn bản, từ đó cải thiện hiệu suất trong các tác vụ phân loại và ứng dụng thực tế.



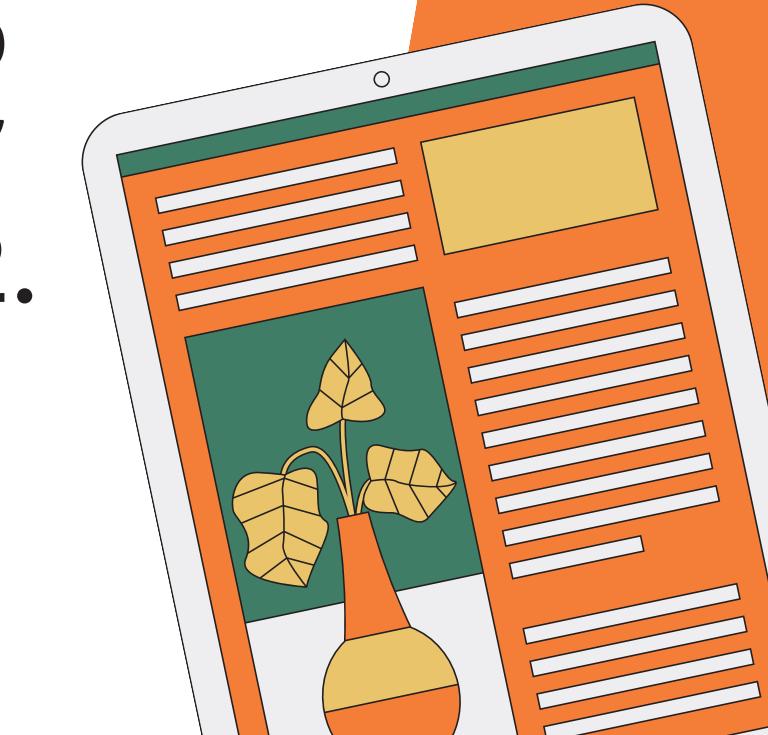
Style	Number of songs
Big Band	362
Blues Contemporary	114
Country Traditional	2065
Dance	2017
Electronica	605
Experimental	733
Folk International	707
Gospel	405
Grunge Emo	302
Hip Hop Rap	801
Jazz Classic	496
Metal Alternative	978
Metal Death	214
Metal Heavy	282
Pop Contemporary	4291
Pop Indie	1147
Pop Latin	838
Punk	113
Reggae	127
RnB Soul	544
Rock Alternative	700
Rock College	977
Rock Contemporary	2890
Rock Hard	2096
Rock Neo Psychedelia	519
Total	24623

folktronica reggae  
 rock dance  
 contemporary ska country adult blues acoustic  
 post-britpop progressive electropop disco indie classic  
 metal electronic jazz hard power neo-medieval pop/rock soft  
 indie alternative/electronic soundtrack euro-pop www  
 heavy hip prog trad american new crossover celtic latin and piano

	All (N = 772.1)	English (N = 521.1)	German (N = 97.9)	French (N = 43.3)	Spanish (N = 27.1)	Japanese (N = 11.0)
1.	Rock	Rock	Pop	Rock	Rock	Rock
2.	Pop	Pop	Rock	Pop	Pop	Pop
3.	Alternative	Alternative	Electronic	Jazz	Jazz	J-Pop
4.	Jazz	Hip-Hop/Rap	Hip-Hop	Hip-Hop	Soundtrack	R&B
5.	Hip-Hop	Hip-Hop	Jazz	Reggae	Latin	Soundtrack
6.	Hip-Hop/Rap	R&B	Alternative	R&B	Dance	Jazz
7.	Soundtrack	Soundtrack	Dance	Soundtrack	House	Electronica/Dance
8.	R&B	Jazz	R&B	Blues	Otros	ロック(Rock)
9.	Electronic	Country	Rock/Pop	Electronic	Blues	Altern. & Punk
10.	Country	Altern. & Punk	Soundtrack	Rap	Electronica	Hip-Hop/Rap



Để đánh giá hiệu suất của các mô hình trên văn bản tiếng Anh và tiếng Slovenia, nghiên cứu sử dụng mẫu ngẫu nhiên từ hai tập dữ liệu được gán nhãn thủ công: EN-GINCO và GINCO. Các mô hình được kiểm tra trên 100 mẫu từ mỗi tập dữ liệu. Vì trình phân loại X-GENRE được huấn luyện trên phần tập huấn luyện của GINCO, các mẫu kiểm tra được lấy từ phần tập kiểm tra. Trong khi đó, tập dữ liệu EN-GINCO không phải là một phần của dữ liệu huấn luyện để tinh chỉnh mô hình X-GENRE.

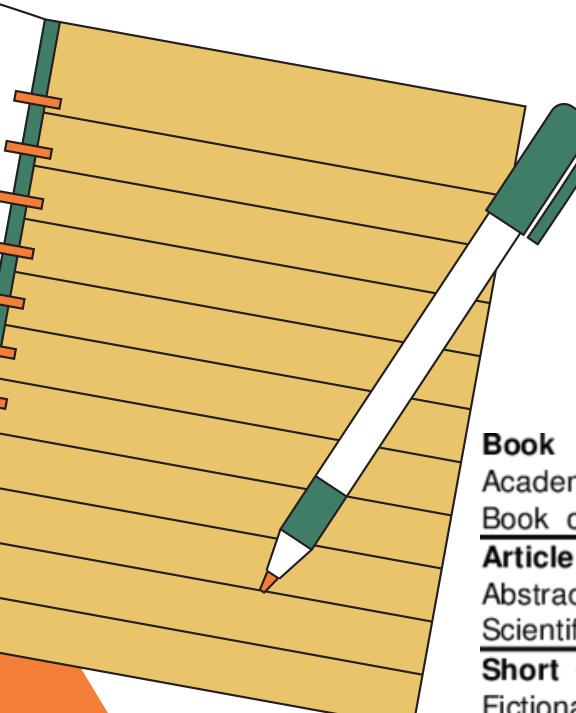


Riêng tập dữ liệu GINCO bao gồm văn bản web tiếng Slovenia từ hai bộ sưu tập văn bản web tiếng Slovenia, slWaC 2.0 và MaCoCu-sl 1.0. Tập dữ liệu EN-GINCO là mẫu văn bản tiếng Anh từ bộ sưu tập văn bản web tiếng Anh enTenTen202. Tập dữ liệu EN-GINCO chưa được công bố trực tuyến, vì vậy, không thể ChatGPT đã thấy nó trước đây. Nó đã được công bố vào năm 2021.

Các mẫu được gán nhãn thủ công với 24 thể loại và hai người gán nhãn có nền tảng ngôn ngữ học tham gia. Độ đồng nhất giữa các người gán nhãn đạt chỉ số Krippendorff's alpha là 0,71, cho thấy độ tin cậy tương đối của dữ liệu được gán nhãn và khó khăn trong việc gán nhãn thủ công cho tác vụ này.



Ở đây, ta sẽ nói thêm về slWaC 2.0, MaCoCu-sl 1.0 và enTenTen202. Nó bao gồm hàng triệu từ và được sử dụng trong nghiên cứu và phát triển ứng dụng xử lý ngôn ngữ tự nhiên (NLP) cho tiếng Slovenia và tiếng Anh.



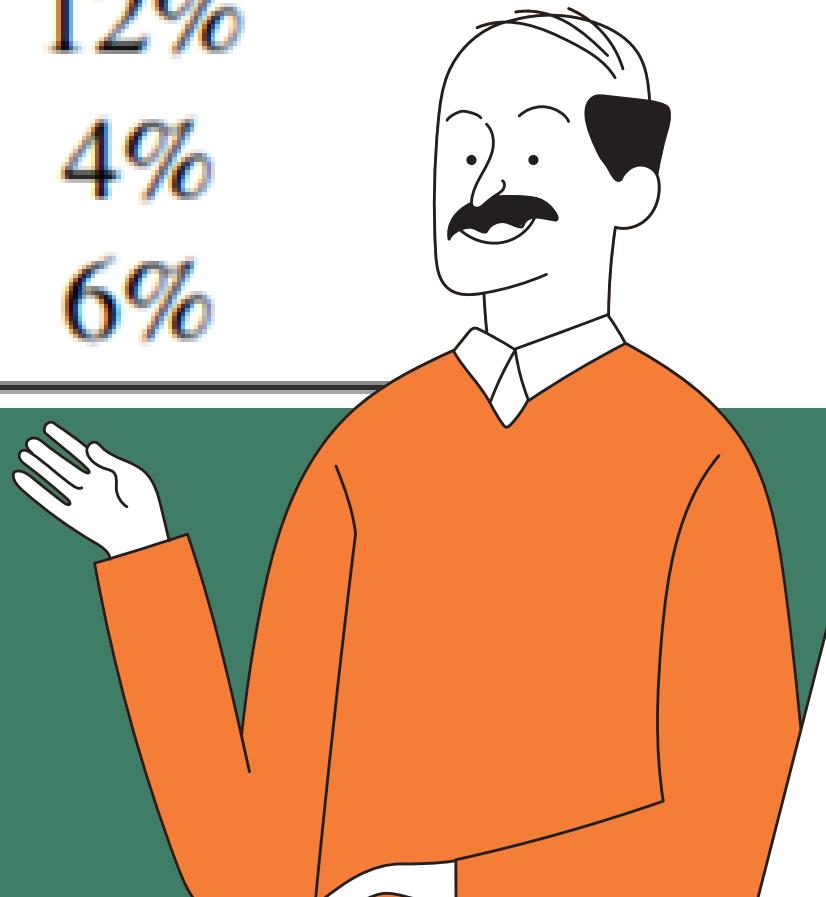
**a. Genre schema (lược đồ thể loại):** là một hệ thống phân loại các thể loại văn bản dựa trên các đặc điểm chung, mục đích của tác giả, chức năng và hình thức thông thường của văn bản. Genre schema giúp nhận dạng và phân loại các văn bản theo thể loại, cho phép các công cụ tìm kiếm thông tin và ứng dụng xử lý ngôn ngữ tự nhiên (NLP) có kết quả tìm kiếm chính xác hơn và hiệu quả hơn trong việc xử lý văn bản.

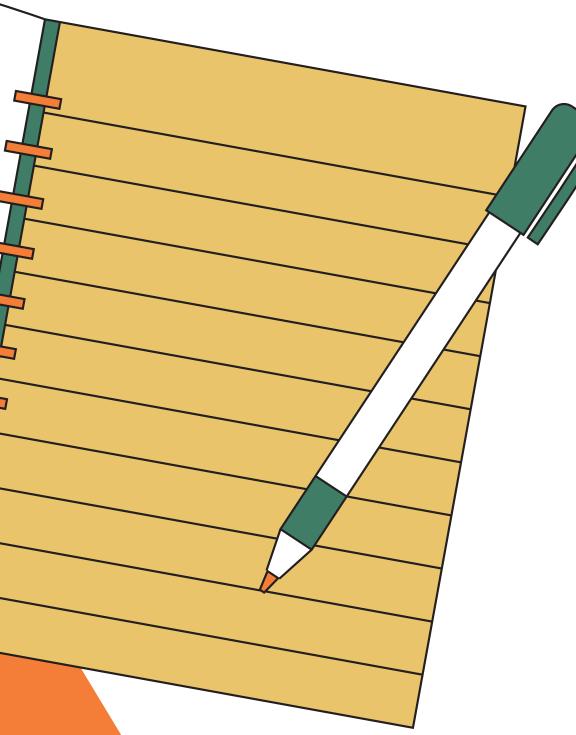
Book	Book of Poetry (4)	Other Book (6)
Academic Monograph (2)		
Book of Fiction (3)	Handbook (5)	
Article		
Abstract (8)	Other Research (10)	News Report (12)
Scientific Article (9)	Magazine Article (11)	
Short Composition		
Fictional Piece (14)	Dramatic Script (16)	Short Biographical Sketch (18)
Poems (15)	Essay (17)	Review (19)
Serial		
Periodicals (News, Mag) (21)	Conference Proceeding (23)	
Journals (22)	Newsletter (24)	
Correspondence		
Email (26)	Memo (29)	
Letter (27)	Telegram (30)	
Treatise		
Thesis (32)	Technical Report (34)	Technical Manual (36)
Business/Operational Rept (33)	Miscellaneous Report (35)	
Information Structure		
List (38)	Table (41)	Programme (44)
Catalogue (39)	Menu (42)	Questionnaire (45)
Raw Data (40)	Form (43)	FAQ (46)
Evidential Document		
Minutes (48)	Financial Record (50)	Slip (52)
Legal Proceedings (49)	Receipt (51)	Contract (53)
Visual Document		
Artwork (55)	Graph (58)	Poster (61)
Card (56)	Diagram (59)	Comics (62)
Chart (57)	Sheet Music (60)	
Other Functional Document		
Guideline (64)	Product Description (70)	Forum Discussion (76)
Regulations (65)	Advertisement (71)	Interview (77)
Manual (66)	Announcement (72)	Notice (78)
Grant/Project Proposal (67)	Appeal/Propaganda (73)	Resume/ CV (79)
Legal Proposal/Order (68)	Exam or Worksheet (74)	Slides (80)
Job/Course/Project Desc. (69)	Factsheet (75)	Speech Transcript (81)

Trong các nghiên cứu và ứng dụng NLP, genre schema thường được sử dụng để gán nhãn thể loại cho các văn bản trong các bộ dữ liệu, giúp huấn luyện và đánh giá các mô hình phân loại văn bản tự động theo thể loại. Genre schema có thể bao gồm nhiều thể loại khác nhau như tin tức, pháp lý, quảng cáo, ý kiến, hướng dẫn và nhiều thể loại khác.

<b>Labels</b>	<b>EN-GINCO</b>	<b>GINCO</b>	<b>X-GENRE training</b>
<b>Information/Explanation</b>	<b>25%</b>	<b>24%</b>	<b>17%</b>
<b>Promotion</b>	<b>22%</b>	<b>17%</b>	<b>16%</b>
<b>Opinion/Argumentation</b>	<b>18%</b>	<b>11%</b>	<b>14%</b>
<b>News</b>	<b>18%</b>	<b>29%</b>	<b>19%</b>
<b>Other</b>	<b>6%</b>	<b>7%</b>	<b>4%</b>
<b>Forum</b>	<b>6%</b>	<b>5%</b>	<b>8%</b>
<b>Instruction</b>	<b>5%</b>	<b>5%</b>	<b>12%</b>
<b>Legal</b>	<b>0%</b>	<b>1%</b>	<b>4%</b>
<b>Prose/Lyrical</b>	<b>0%</b>	<b>1%</b>	<b>6%</b>

Trong thử nghiệm, tác giả bài báo sử dụng lược đồ thể loại X-GENRE, một sự tổng hợp của nhiều lược đồ được áp dụng cho các tập dữ liệu khác nhau như CORE, FTD và GINCO. Mục đích của lược đồ này là thân thiện với người dùng hơn so với các lược đồ cụ thể khác và cho phép kết hợp dữ liệu huấn luyện từ các tập dữ liệu khác nhau, tạo ra mô hình ổn định hơn. Tác giả đã ánh xạ các nhãn gốc trong cả hai tập dữ liệu kiểm tra, được gắn nhãn với một lược đồ có độ chi tiết hơn, sang lược đồ X-GENRE. Lược đồ cuối cùng bao gồm 9 nhãn như hình 1 ở trên.





## 5. Các mô hình:

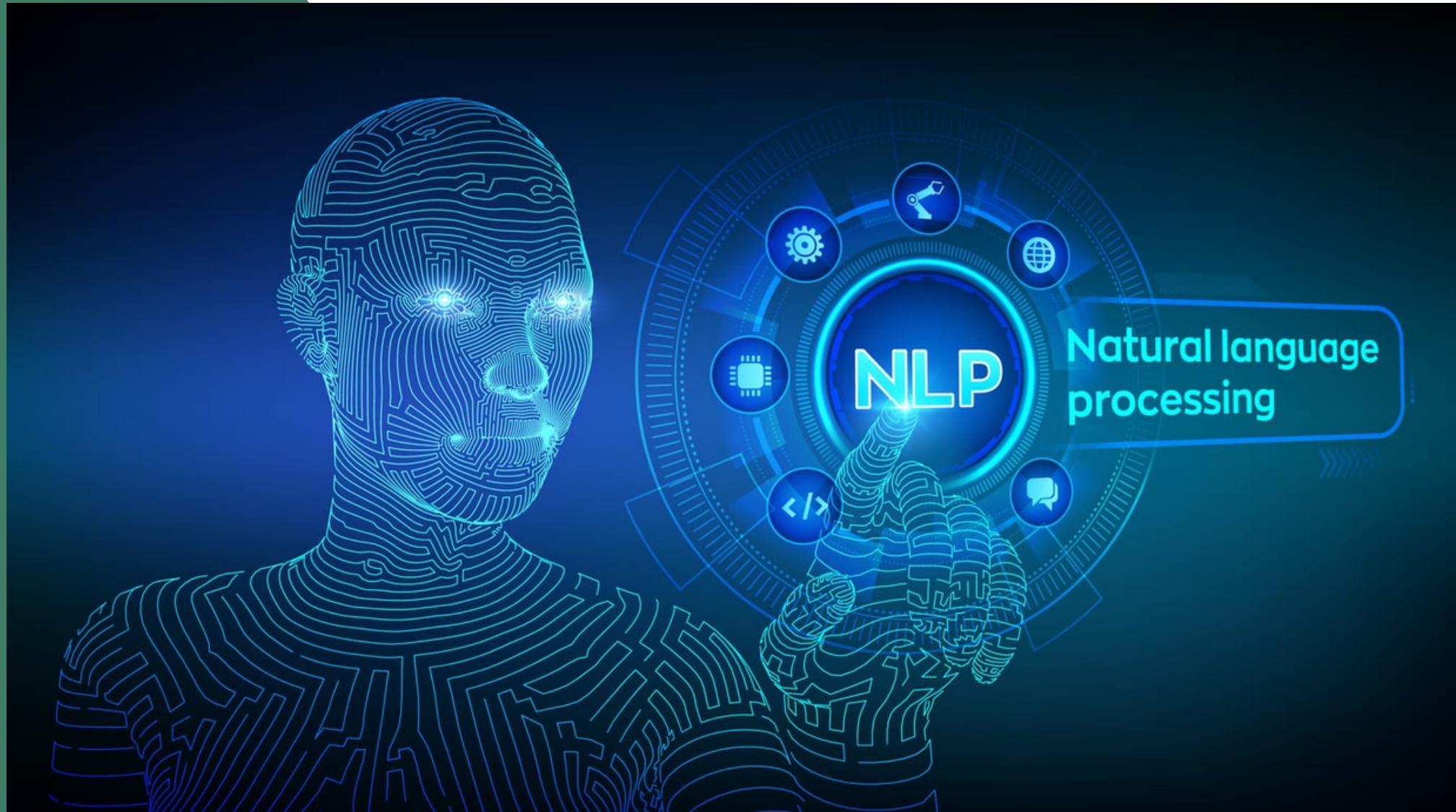


# A. ChatGPT:

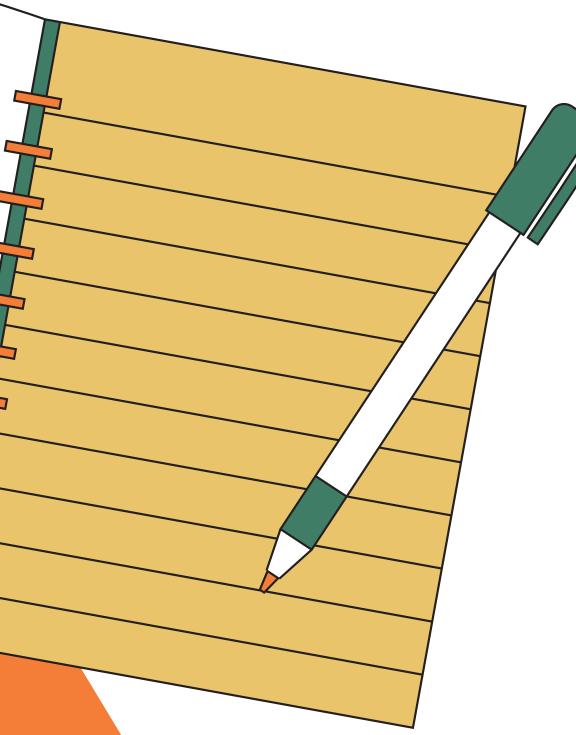
ChatGPT là một mô hình ngôn ngữ lớn do OpenAI cung cấp, được tinh chỉnh trên mô hình GPT-3.5 (OpenAI, 2023) và được tối ưu hóa cho điện thoại bằng phương pháp học tăng cường với phản hồi từ con người. Nói cách khác, mô hình được đào tạo để tạo ra các câu trả lời tốt nhất dựa trên đánh giá từ con người. Các tác giả đã sử dụng phiên bản ChatGPT 13/2 và thực hiện các thí nghiệm trong khoảng thời gian từ ngày 24/2 đến 2/3/2023.



## B. X-Genre:



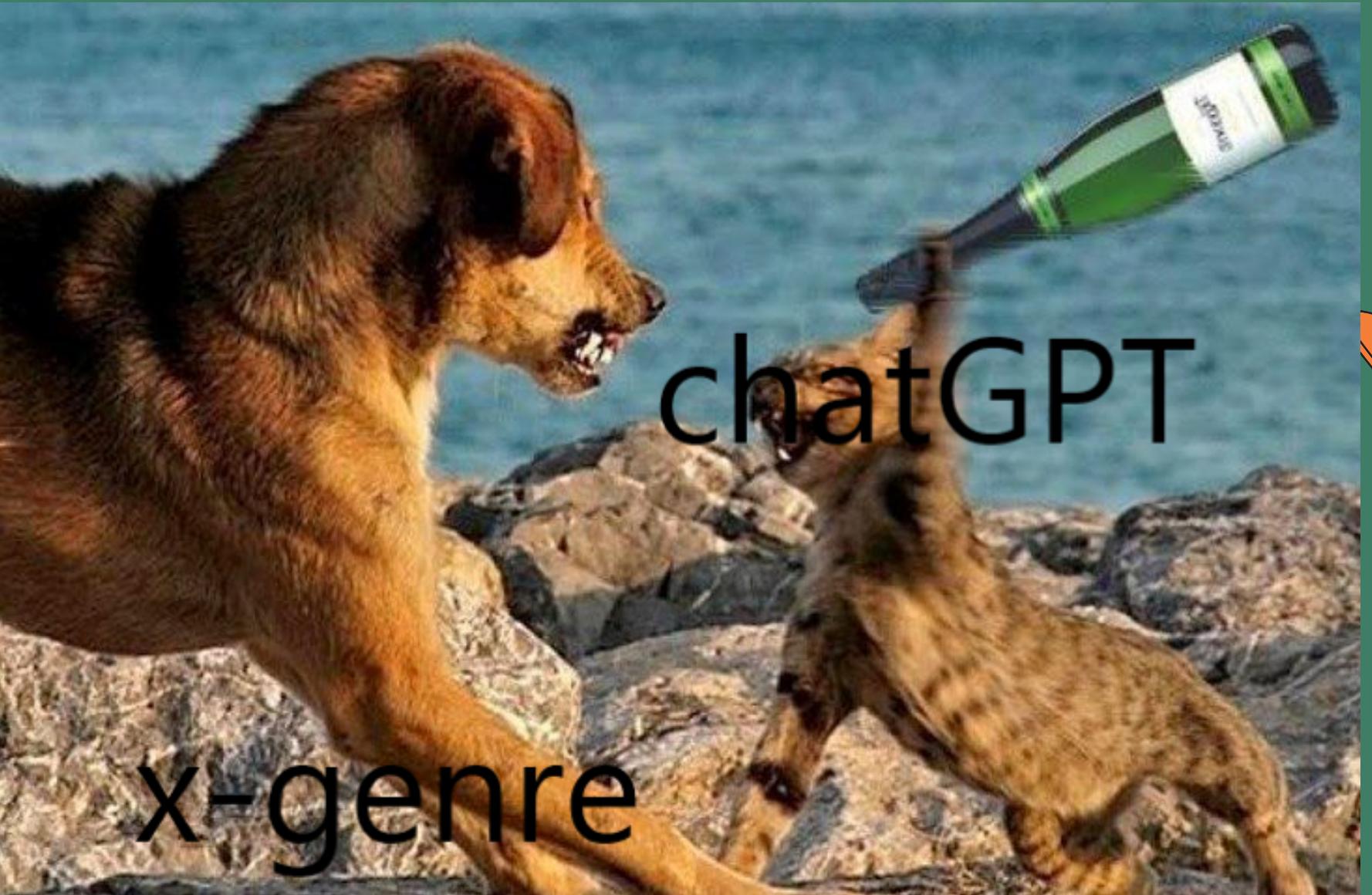
Ta so sánh mô hình ChatGPT với một mô hình dựa trên XLM-RoBERTa được tinh chỉnh trên các tập dữ liệu được chú thích theo thể loại (genre). Mô hình X-GENRE được tinh chỉnh trên khoảng 1.700 trường hợp từ ba bộ dữ liệu với nhãn thể loại được chú thích bằng tay: CORE (Egbert et al., 2015), FTD (Sharoff, 2018) và GINCO (Kuzman et al., 2022b). Mô hình này đạt điểm F1 micro và macro giữa 0,79 và 0,80 trong kịch bản kiểm tra trong tập dữ liệu. Khi so sánh với các mô hình chỉ được đào tạo trên một bộ dữ liệu, kết quả cho thấy mô hình X-GENRE vượt trội hơn. Mô hình này có sẵn miễn phí trên kho lưu trữ Hugging Face.



# **6. Lời nhắc ChatGPT và phỏng đoán các thể loại:**



Để phân loại văn bản theo thể loại với ChatGPT, nhóm tác giả sử dụng các lời nhắc trên nền tảng OpenAI và trích xuất thủ công các danh mục và giải thích từ các câu trả lời của nó. Trong lời nhắc, tác giả xác định các tiêu chí chính để xác định thể loại và danh mục mà mô hình có thể lựa chọn. tác giả cũng yêu cầu mô hình cung cấp lý do giải thích lựa chọn của mình và cung cấp văn bản cần phân loại. Lời nhắc được lặp lại với mỗi văn bản. tác giả sử dụng các lớp thể loại được sử dụng bởi bộ phân loại X-GENRE để có thể so sánh hai mô hình.



Một ví dụ về lời nhǎc tiếng Anh và câu trả lời của ChatGPT gồm nhǎn và giải thích. Trong hầu hết trường hợp, câu trả lời của ChatGPT tuân theo dạng của ví dụ, bao gồm nhǎn trong một dòng và giải thích trong đoạn văn tiếp theo. Tuy nhiên, trong một số trường hợp, nhǎn là một phần của giải thích. Lời nhǎc tiếng Slovenia là bản dịch của lời nhǎc tiếng Anh, trong đó cả tên của các lớp cũng được dịch sang tiếng Slovenia. Khi sử dụng lời nhǎc tiếng Slovenia, câu trả lời của ChatGPT cũng bằng tiếng Slovenia.

## Ví dụ của tác giả về tiếng Anh



Please classify the following text according to genre (defined by function of the text, author's purpose and form of the text) and explain your decision. You can choose from the following classes: News, Legal, Promotion, Opinion/Argumentation, Instruction, Information/Explanation, Prose/Lyrical, Forum, Other.

The text to classify: Shower pods install in no time. . .

- <p> 1. Prepare the floor with the waste and the water supply pipes.
- <p> 2. Attach shower equipment to the shower pod shell running flexible tails (H&C or just C) down back.
- <p> 3. Move unit into position connecting water supplies on the way and the waste outlet trap.
- <p> 4. Having secured the shower pod shell to the building structure doors may now be fitted.

Example of the ChatGPT's answer:

Genre: Instruction

Explanation: The text provides step-by-step instructions for installing shower pods. It is structured in a sequential manner, detailing the process of preparing the floor, attaching shower equipment, moving the unit into position, and securing the shower pod shell. The language used is imperative, giving clear and concise commands to the reader.

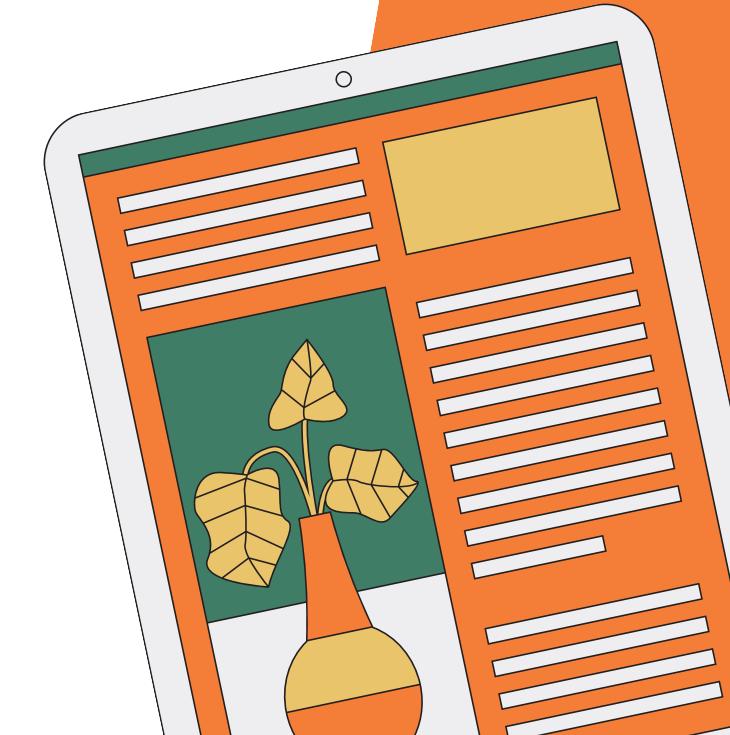
There is no attempt to persuade, promote, or argue for any particular point of view.

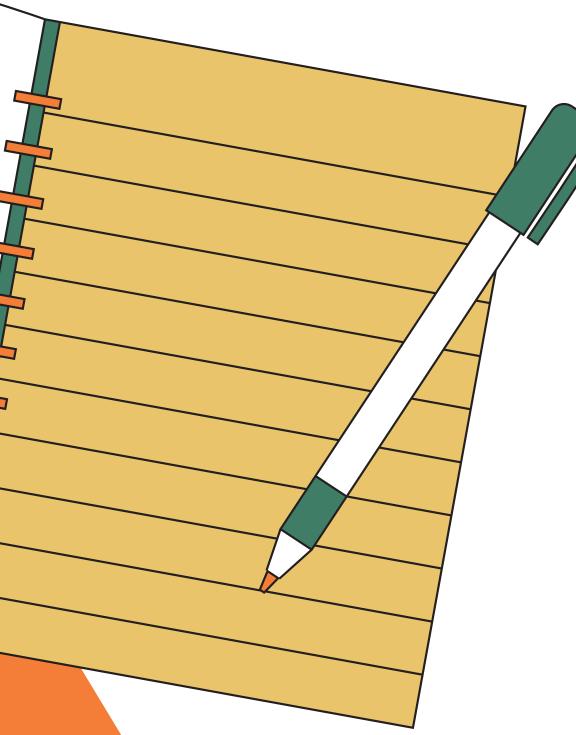
Instead, the purpose of the text is to instruct the reader on how to install shower pods.

Therefore, the text can be classified as an instruction manual or guide.



Trong một số trường hợp, đầu ra của ChatGPT gồm hai lớp thay vì một. Nếu một trong số các lớp phù hợp với nhãn thực tế, tác giả sử dụng lớp này để đánh giá. Ngoài ra, trong một số trường hợp, ChatGPT trả lời bằng nhãn mới không thuộc tập hợp lớp thể loại, chẳng hạn như "Đánh giá" và "Chia buồn". tác giả đã thử nghiệm yêu cầu nó sửa đổi quyết định và chọn nhãn từ tập hợp đóng.





# **7. So sánh với mô hình tinh chỉnh <fine-tuned model>:**

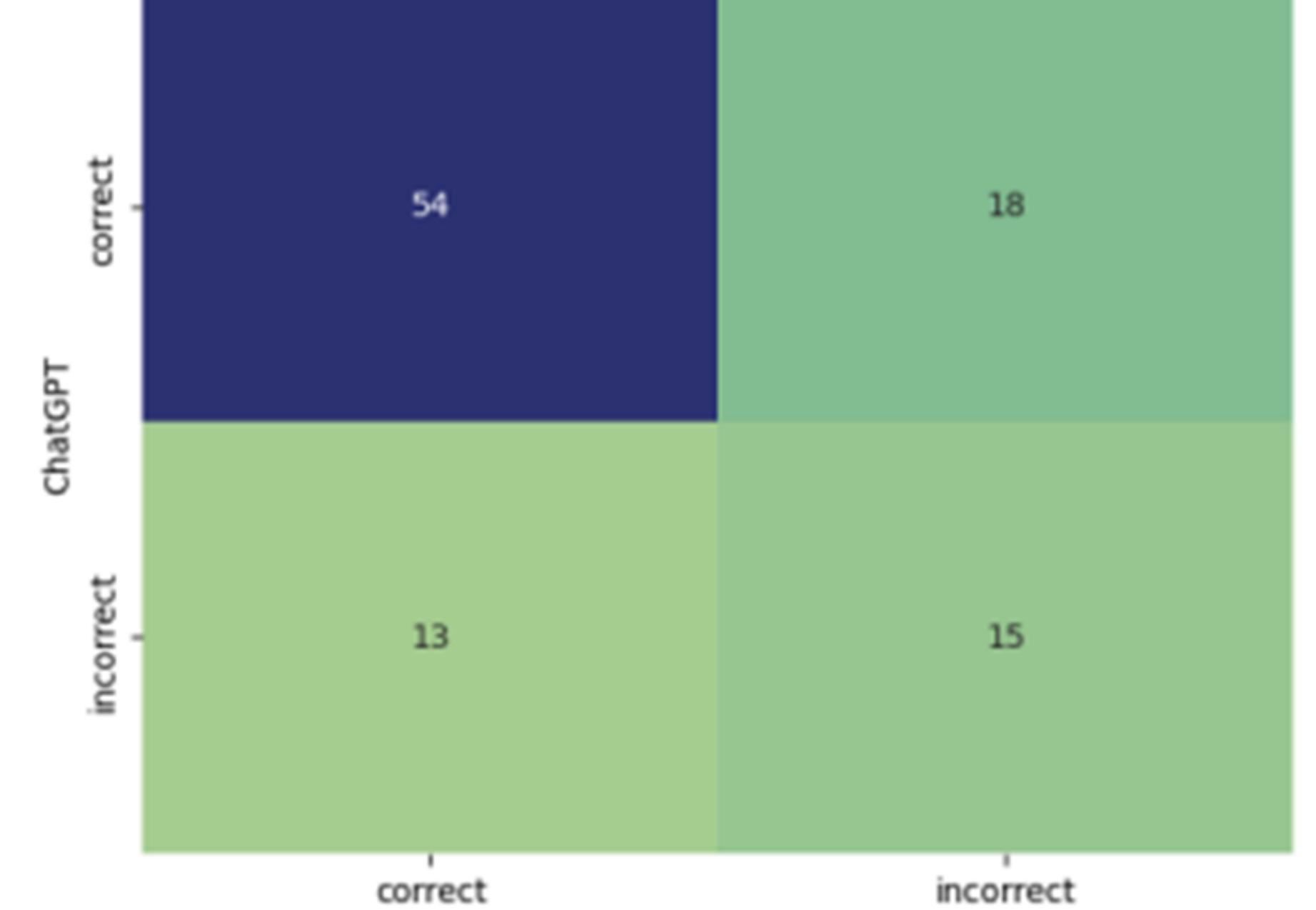


Tác giả so sánh hai mô hình trong ba kịch bản:  
Trên bộ kiểm tra tiếng Anh (EN-GINCO) với lời  
nhắc bằng tiếng Anh, trên bộ kiểm tra tiếng  
Slovenia (GINCO) với lời nhắc bằng tiếng Anh  
và trên bộ kiểm tra tiếng Slovenia (GINCO) với  
dấu nhắc tiếng Slovenia. Trong hai trường hợp  
sau, chỉ có ngôn ngữ của lời nhắc là khác,  
trong khi các trường hợp văn bản để phân  
loại là giống nhau. Các kết quả được thể hiện  
trong Bảng 2.

Bảng 2: So sánh ChatGPT và mô hình X-GENRE  
tinh chỉnh trên hai bộ thử nghiệm và với lời  
nhắc trong hai bộ ngôn ngữ.

Mô hình 1: So sánh sự khác biệt về dự đoán  
đúng và sai giữa

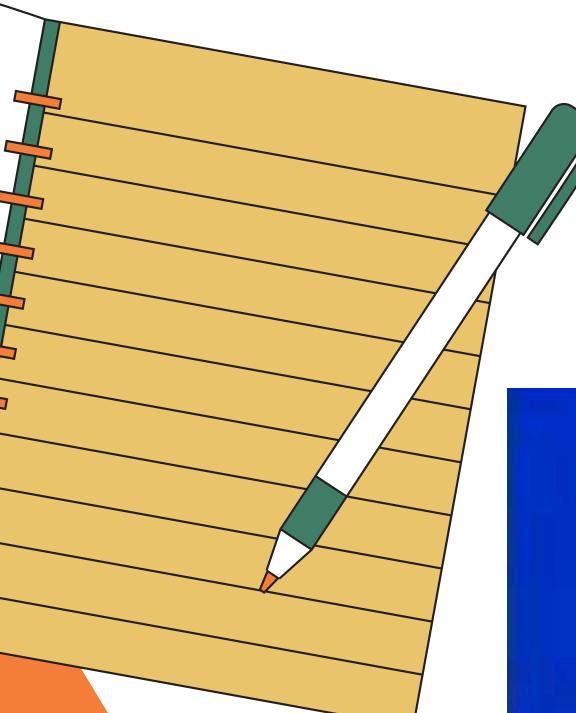
ChatGPT và X-GENRE.



Test set	Prompt	Model	Micro F1	Macro F1	Accuracy
EN-GINCO	EN	ChatGPT	0.74	0.66	0.72
		X-GENRE	0.67	0.61	0.67
SL-GINCO	EN	ChatGPT	0.75	0.64	0.75
		X-GENRE	0.91	0.91	0.91
	SL	ChatGPT	0.68	0.56	0.68
		X-GENRE	0.91	0.91	0.91

# 8. Kết luận - Hướng phát triển mới:





**THANKS FOR LISTENING**

CONTACT: [dat20026969@gmail.com](mailto:dat20026969@gmail.com)

