Week 3 - Machine Learning for NLP

September 13, 2025

1 Các bài toán cơ bản trong NLP

1.1 Classification

- **Text Classification:** Gán một hoặc nhiều nhãn cho một câu hoặc đoạn văn bản. *Ví dụ:* Phân loại email thành `spam` hoặc `không spam`; phân loại bài báo thành các chủ đề `thể thao`, `giáo dục`, `chính trị`.
- **Sentiment Analysis:** Xác định cực tính (tích cực, tiêu cực, trung lập) của văn bản. *Ví dụ:* "Dịch vụ ở quán này rất tốt" → Tích cực. "Dịch vụ ở quán này tệ" → Tiêu cực. "Dịch vụ ở quán này không tệ" → Trung lập.

1.2 Information Retrieval and Document Re-ranking

- **Sentence/Document Similarity:** Xác định mức độ tương đồng giữa hai văn bản. *Ví dụ:* Tìm tài liệu liên quan khi gõ từ khóa "Xử lý ngôn ngữ tự nhiên".
- **Question Answering (QA):** Trả lời câu hỏi bằng ngôn ngữ tự nhiên. *Ví dụ:* Câu hỏi: "Ngày Quốc khánh Việt Nam là ngày nào?" → Đáp án: "2/9".

1.3 Text-to-Text Generation

- **Machine Translation:** Dịch từ ngôn ngữ này sang ngôn ngữ khác. *Ví dụ:* "Good morning" → "Chào buổi sáng".
- **Text Summarization:** Tạo bản tóm tắt văn bản nhưng vẫn giữ được ý nghĩa. *Ví dụ:* Văn bản: "AI trong nông nghiệp giúp dự báo mùa vụ, phân tích dữ liệu cảm biến và hình ảnh vệ tinh để khuyến nghị gieo trồng, tưới tiêu, phòng trừ sâu1bệnh." → Tóm tắt: "AI giúp dự báo mùa vụ và tối ưu gieo trồng, tưới tiêu, phòng trừ sâu bệnh".
- **Text Simplification:** Làm văn bản dễ đọc, dễ hiểu hơn. *Ví dụ:* "Quang hợp là quá trình cây xanh dùng ánh sáng để tổng hợp carbohydrate từ CO_2 và nước." → "Quang hợp là quá trình cây xanh dùng ánh sáng để tạo thức ăn từ CO_2 và nước".

1.5 Topics and Keywords

- **Topic Modeling:** Xác định các chủ đề tiềm ẩn trong tập văn bản. *Ví dụ:* Diễn đàn bóng đá → Chủ đề: `Thể thao`.
- **Keyword Extraction:** Trích xuất từ khóa quan trọng. *Ví dụ:* Văn bản về "AI trong nông nghiệp" → Từ khóa: `AI`, `nông nghiệp`, `mùa vụ`.

1.6 Text Reasoning

- **Common Sense Reasoning:** Suy luận dựa trên kiến thức thường thức. *Ví dụ:* "Nếu trời mưa, mặt đất sẽ như thế nào?" → "Ướt".
- **Natural Language Inference (NLI):** Xác định mối quan hệ giữa tiền đề và giả thuyết. *Ví dụ:* Tiền đề: "Một người đang chơi guitar" Giả thuyết: "Có ai đó đang chơi nhạc" → Bao hàm.

1.7 Text Preprocessing

- **Coreference Resolution:** Liên kết các đề cập cùng thực thể. *Ví dụ:* "Obama phát biểu. Ông ấy nói ..." → "Ông ấy" = "Obama".
- **POS Tagging:** Gán nhãn từ loại. *Ví dụ:* "Học sinh /N học /V chăm chỉ /Adj".
- **Word Sense Disambiguation:** Xác định nghĩa đúng của từ trong ngữ cảnh. *Ví dụ:* "bank" trong "river bank" = bờ sông, trong "money bank" = ngân hàng.
- **Grammatical Error Correction:** Sửa lỗi ngữ pháp. *Ví dụ:* "Tôi đi chợ ngày hôm qua" → "Tôi đã đi chợ ngày hôm qua".

2 Các bài toán NLP phổ biến với Tiếng Việt

2.1 Sentiment Analysis và Text Classification

Nhiệm vụ phân loại văn bản vào các nhóm định trước như tích cực, tiêu cực, trung lập hoặc theo chủ đề (chính trị, thể thao, công nghệ). *Ví dụ:* "Sản phẩm này thật tuyệt, tôi sẽ mua lại" → Tích cực. "Dịch vụ quá kém, tôi sẽ không quay lại" → Tiêu cực.

Ứng dụng: theo dõi dư luận trên mạng xã hội (Facebook, Zalo) hoặc đánh giá sản phẩm (Shopee reviews). Các bộ dữ liệu benchmark: UIT-VSFC, VLSP Sentiment Analysis Challenge.

2.2 Tokenization/Word Segmentation

Khác với tiếng Anh, nơi khoảng trắng phân tách rõ ràng giữa các từ, trong tiếng Việt khoảng trắng chỉ chia tách *âm tiết*, chứ không phải từ hoàn chỉnh. Điều này khiến bài toán tách từ trở thành một bước tiền xử lý quan trọng.

Ví dụ: Cụm "công nghệ thông tin" nên được phân tách thành [công_nghệ, thông_tin] thay vì [công, nghệ, thông, tin]. Nếu tách sai, mô hình sẽ không hiểu đúng nghĩa, dẫn đến sai lệch trong các tác vụ phía sau như gán nhãn từ loại (POS tagging), phân tích cú pháp, hay dịch máy.

Một ví dụ khác: cụm "xe máy điện" phải được nhận diện thành [xe_máy, điện] để bảo toàn ý nghĩa. Nếu tách thành [xe, máy, điện], hệ thống có thể hiểu nhầm là "chiếc xe, một cái máy, và điện" — hoàn toàn không đúng.

Các công cụ tách từ tiếng Việt phổ biến hiện nay bao gồm **underthesea**, **VnCoreNLP**, và **pyvi**.

2.3 POS Tagging và Named Entity Recognition (NER)

POS Tagging: là quá trình gán nhãn từ loại cho các từ trong câu

Ví dụ: "Học sinh đang đọc sách trong thư viện"→ [Học_sinh/NOUN, đang/ADV, đọc/VERB, sách/NOUN, trong/ADP, thư_viện/NOUN].

Named Entity Recognition (NER): Là bài toán nhận diện và phân loại thực thể trong văn bản thành các nhóm như `Người`, `Địa điểm`, `Tổ chức`, `Thời gian`, v.v.

Ví dụ: "UIT là một trường đại học tại Việt Nam." → [UIT/ORGANIZATION, Việt Nam/LOCATION].

"Trần Đại Quang từng là Chủ tịch nước Việt Nam." → [Trần Đại Quang/PERSON, Việt Nam/LOCATION].

2.4 Dịch máy (Machine Translation)

Dịch Anh–Việt và Việt–Anh là tác vụ quan trọng trong giao tiếp song ngữ, giáo dục và thương mại. *Ví dụ:* English → Vietnamese: "How are you today?" → "Hôm nay bạn thế nào?". Vietnamese → English: "Tôi đang học xử lý ngôn ngữ tự nhiên" → "I am studying natural language processing".

Thách thức: xử lý dấu thanh, từ ghép và dịch đúng thành ngữ văn hoá. *Ví dụ:* "một cây làm chẳng nên non" → "unity makes strength" (tương đương "united we stand, divided we fall").

3 Các bài toán có thể giải quyết được bằng LLM

Các mô hình ngôn ngữ lớn (LLM) đã thể hiện khả năng vượt trội trong nhiều nhiệm vụ Xử lý Ngôn ngữ Tự nhiên (NLP), đặc biệt khi được khai thác thông qua kỹ thuật *prompting*. Người dùng có thể hướng dẫn LLM bằng ngôn ngữ tự nhiên để giải quyết các bài toán khác nhau mà không cần huấn luyện lại mô hình.

3.1 Phân loại văn bản (Text Classification)

Ví dụ prompt:

```
Hãy phân loại câu sau thành tích cực, tiêu cực hoặc trung
lập: "Bộ phim này thật sự rất hay và cảm động."
```

Kết quả mong đợi: *Tích cực.*

3.2 Phân tích cảm xúc (Sentiment Analysis)

Ví dụ prompt:

```
Hãy xác định cảm xúc chính trong câu: "Tôi vừa mất ví và
cảm thấy vô cùng thất vọng."
```

Kết quả mong đợi: *Tiêu cực.*

3.3 Question Answering

Ví dụ prompt:

```
Dựa trên đoạn văn sau, hãy trả lời câu hỏi: "Đoạn văn:
Isaac Newton phát minh ra giải tích và định luật vạn vật
hấp dẫn. Câu hỏi: Ai là người phát minh ra giải tích?"
```

Kết quả mong đợi: *Isaac Newton.*

3.4 Dịch máy (Machine Translation)

Ví dụ prompt:

```
Dịch câu sau từ tiếng Việt sang tiếng Anh: "Học sinh đang
chơi bóng đá ở sân trường."
```

Kết quả mong đợi: *The students are playing football in the schoolyard.*

3.5 Sinh văn bản (Text Generation)

Ví dụ prompt:

```
Hãy viết một đoạn văn ngắn (2 câu) giới thiệu về thành
phố Hà Nội.
```

Kết quả mong đợi: *Hà Nội là thủ đô của Việt Nam với hơn một nghìn năm lịch sử. Thành phố nổi tiếng với Hồ Hoàn Kiếm, Phố Cổ và nền ẩm thực phong phú.*

3.6 Tóm tắt văn bản (Text Summarization)

Mục tiêu là tạo phiên bản rút gọn nhưng vẫn giữ được ý chính.

Ví dụ prompt:

```
Hãy tóm tắt đoạn văn sau trong một câu: "Internet đã thay
đổi cách con người làm việc, học tập và giao tiếp. Ngày
nay, chúng ta có thể dễ dàng kết nối với nhau thông qua
email, mạng xã hội và các ứng dụng trò chuyện trực tuyến."
```

Kết quả mong đợi: *Internet đã cách mạng hóa việc học tập, làm việc và giao tiếp của con người.*

3.7 Sinh câu đồng nghĩa (Paraphrase Generation)

Ví dụ prompt:

```
Viết lại câu sau với ý nghĩa tương tự: "Tôi rất vui khi
được gặp lại bạn."
```

Kết quả mong đợi: *Thật hạnh phúc khi tôi có thể gặp bạn lần nữa.*

3.8 Trích xuất quan hệ (Relation Extraction)

Ví dụ prompt:

```
Hãy tìm quan hệ giữa các thực thể trong câu: "Albert Einstein
sinh ra ở Đức và làm việc tại Mỹ."
```

Kết quả mong đợi:

- (Albert Einstein, sinh ra, Đức)
- (Albert Einstein, làm việc tại, Mỹ)

3.9 Named Entity Recognition - NER)

Ví dụ prompt: `Hãy nhận dạng thực thể trong câu sau: "Microsoft hợp tác với Đại học Stanford` **Kết quả mong đợi:**

- Microsoft → Organization
- Đại học Stanford → Organization

4 Các bài toán NLP ứng dụng để xây dựng Chatbot

4.1 Intent Recognition (Text Classification)

Xác định mục đích của người dùng trong một phát ngôn, ví dụ: đặt vé, kiểm tra thời tiết, hay hỏi thông tin sản phẩm. *Ví dụ:* "Tôi muốn đặt vé máy bay đi Đà Nẵng" → Intent: `Đặt vé`.

4.2 Named Entity Recognition (NER)

Nhận diện các thực thể trong lời nói của người dùng như tên, ngày, địa điểm hoặc tên sản phẩm. *Ví dụ:* "Đặt vé đi Hà Nội vào ngày mai" → {Hà Nội: `Location`, ngày mai: `Date`}.

4.3 Dialogue Generation

Sinh phản hồi tự nhiên và phù hợp với ngữ cảnh, giúp chatbot duy trì hội thoại nhiều lượt một cách mạch lạc. *Ví dụ:* Người dùng: "Tôi muốn đặt bàn cho tối nay." Chatbot: "Anh/chị muốn đặt bàn cho bao nhiêu người ạ?"

4.4 Question Answering and Retrieval

Cung cấp câu trả lời chính xác dựa trên tri thức có sẵn hoặc từ nguồn bên ngoài. *Ví dụ:* Người dùng: "Thủ đô của Việt Nam là gì?" Chatbot: "Thủ đô của Việt Nam là Hà Nội."

Việc tích hợp **retrieval-augmented generation (RAG)** có thể giúp cải thiện độ chính xác và giảm thiểu hiện tượng *hallucination*.

5 Các thư viện có sẵn để giải quyết các bài toán NLP

Trong quá trình xây dựng hệ thống xử lý ngôn ngữ tự nhiên (NLP), nhiều thư viện và framework đã được phát triển, hỗ trợ từ bước tiền xử lý văn bản cho đến huấn luyện và triển khai mô hình hiện đại. Dưới đây là một số nhóm thư viện phổ biến:

5.1 Nền tảng cơ bản

- **SpaCy, NLTK:** Cung cấp các công cụ xử lý văn bản truyền thống như tokenization, stemming, lemmatization, POS tagging, parsing và các thuật toán cơ bản cho NLP.
- **underthesea:** Thư viện chuyên biệt cho tiếng Việt, hỗ trợ tokenization, gán nhãn từ loại (POS tagging), nhận diện thực thể tên (NER), và phân tích quan hệ cú pháp (dependency parsing). Rất hữu ích khi xử lý ngôn ngữ tiếng Việt.

5.2 Machine Learning truyền thống

- **scikit-learn:** Hỗ trợ xây dựng baseline với các mô hình học máy kinh điển như Naive Bayes, SVM, Logistic Regression. Thư viện này cũng cung cấp các công cụ trích xuất đặc trưng văn bản (Bag-of-Words, TF-IDF) và đánh giá mô hình.

5.3 Mô hình học sâu và pretrained

- **HuggingFace Transformers:** Kho thư viện chứa hàng trăm mô hình ngôn ngữ tiền huấn luyện (pretrained) như BERT, RoBERTa, T5, BART, GPT variants, mBERT, XLM-R, v.v. Hỗ trợ fine-tuning cho hầu hết các tác vụ NLP: phân loại văn bản, NER, dịch máy, sinh văn bản, tóm tắt văn bản, và QA. Đồng thời, thư viện này còn tích hợp với PyTorch, TensorFlow, dễ dàng sử dụng trong nghiên cứu và sản phẩm thực tế.

5.4 Thư viện chuyên biệt

- **Whisper:** Mô hình ASR (Automatic Speech Recognition) mạnh mẽ của OpenAI, hỗ trợ nhận diện giọng nói và chuyển đổi giọng nói thành văn bản (speech-to-text) đa ngôn ngữ, trong đó có tiếng Việt.

- **Rasa:** Framework mã nguồn mở cho phát triển chatbot, cung cấp các module cho nhận diện ý định (intent detection), đslot filling), và dialogue management.
- **Haystack:** Thư viện dành cho xây dựng hệ thống QA (Question Answering) và Information Retrieval. Hỗ trợ pipeline kiểu retriever-reader, tích hợp với các công cụ tìm kiếm và cơ sở dữ liệu vector như Elasticsearch, Milvus.

6 Bài tập ứng dụng

Đầu tiên, sử dụng thư viện hỗ trợ tiếng Việt **underthesea** để minh họa hai bài toán NLP là Sentiment Analysis và Named Entity Recognition. Sau đó, kết hợp tiền xử lý văn bản với mô hình học máy cho bài toán Text Classification.

6.1 Minh họa Sentiment Analysis

```
Sản phẩm này thật tuyệt vời! --> positive
Tôi rất hài lòng về chất lượng dịch vụ. --> positive
Thức ăn ngon, nhân viên phục vụ nhiệt tình. --> positive
Trải nghiệm thật sự đáng nhớ, tôi sẽ quay lại. --> positive
Giá rẻ mà chất lượng tốt bất ngờ. --> positive
Dịch vụ ở đây quá tệ. --> negative
Tôi thất vọng về sản phẩm này. --> negative
Giao hàng chậm, nhân viên thô lỗ. --> negative
Chất lượng quá kém, không đáng tiền. --> negative
Tôi sẽ không bao giờ quay lại nữa. --> negative
Hôm nay trời nắng. --> positive
Tôi đi siêu thị mua ít đồ. --> negative
Cuốn sách này có 300 trang. --> None
Anh ấy sống ở Hà Nội. --> None
Tôi có lịch họp vào lúc 9 giờ sáng. --> None
Dịch vụ ở quán này không tệ. --> negative
Sản phẩm này không hề rẻ chút nào. --> positive
Tôi không ghét món ăn này. --> negative
Bộ phim không phải dở, nhưng cũng không hay. --> negative
Thái độ của nhân viên không tốt lắm. --> negative
```

Hình 1: Kết quả phân tích cảm xúc với underthesea

Ở một vài ví dụ, underthesea thực hiện khá tốt với những câu đơn giản, chẳng hạn: "Sản phẩm này thật tuyệt vời" → *positive*.

Tuy nhiên, với những câu phức tạp hơn như: "Sản phẩm này không hề rẻ chút nào" kết quả mong đợi là *negative*, nhưng hệ thống lại dự đoán *positive*.

Điều này cho thấy thư viện vẫn còn hạn chế trong việc xử lý ngữ nghĩa phủ định phức tạp.

6.2 Minh họa Named Entity Recognition

```
Input: Ông Tô Lâm là Tổng bí thư Đảng Cộng sản Việt Nam.
NER: [('Ông', 'Nc', 'B-NP', 'O'), ('Tô Lâm', 'Np', 'B-NP', 'B-PER'), ('là', 'V', 'B-VP', 'O'), ('Tổng bí thư',
'N', 'B-NP', 'O'), ('Đảng Cộng sản Việt Nam', 'V', 'B-VP', 'O'), ('.', 'CH', 'O', 'O')]
--------------------------------------------------------------------------------
Input: Lan làm việc tại Công ty FPT ở Hà Nội.
NER: [('Lan', 'Np', 'B-NP', 'B-PER'), ('làm việc', 'V', 'B-VP', 'O'), ('tại', 'E', 'B-PP', 'O'), ('Công ty', 'N
p', 'B-NP', 'B-LOC'), ('FPT', 'Np', 'I-NP', 'I-LOC'), ('ở', 'E', 'B-PP', 'O'), ('Hà Nội', 'Np', 'B-NP', 'B-LOC'),
('.', 'CH', 'O', 'O')]
--------------------------------------------------------------------------------
```

Hình 2: Kết quả nhận dạng thực thể có tên với underthesea

Kết quả minh họa cho thấy **underthesea** hoạt động khá tốt với các ví dụ cơ bản.

6.3 Text Classification

Ở phần này, tiến hành tiền xử lý văn bản và sử dụng mô hình dự đoán.

Accuracy: 0.7007

	precision	recall	f1-score	support
0	0.69	0.72	0.70	221
1	0.62	0.55	0.58	177
2	0.78	0.80	0.79	210
accuracy			0.70	608
macro avg	0.69	0.69	0.69	608
weighted avg	0.70	0.70	0.70	608

(a) Sau khi loại bỏ stopwords

Accuracy: 0.7780

	precision	recall	f1-score	support
0	0.78	0.86	0.81	221
1	0.68	0.61	0.64	177
2	0.85	0.84	0.85	210
accuracy			0.78	608
macro avg	0.77	0.77	0.77	608
weighted avg	0.78	0.78	0.78	608

(b) Khi giữ lại stopwords

Hình 3: So sánh kết quả khi xử lý stopwords

Tiền xử lý bao gồm:

- Chuyển toàn bộ văn bản về chữ thường (lowercasing)
- Loại bỏ ký tự đặc biệt
- Loại bỏ link và URL
- Loại bỏ emoji và chữ số

Đặc trưng được trích xuất bằng phương pháp TF-IDF, sau đó kết hợp với mô hình SVM để dự đoán. Kết quả thực nghiệm cho thấy việc loại bỏ stopwords hoàn toàn lại làm hiệu suất giảm đáng kể. Điều này có thể là do khi bỏ giữ vai trò quan trọng về mặt ngữ nghĩa trong tiếng Việt (ví dụ:

"không", "chẳng", "chưa"). Đây là những từ dừng trong tiếng Anh, nhưng trong tiếng Việt chúng quyết định tính phủ nên ảnh hưởng trực tiếp đến kết quả phân loại.