ISSN 2354-0575

# SỬ DỤNG BERT CHO TÓM TẮT TRÍCH RÚT VĂN BẢN

**Đỗ Thị Thu Trang, Trịnh Thị Nhị, Ngô Thanh Huyền**
*Trường Đại học Sư phạm Kỹ thuật Hưng Yên*

Ngày tòa soạn nhận được bài báo: 03/03/2020
Ngày phản biện đánh giá và sửa chữa: 15/05/2020
Ngày bài báo được duyệt đăng: 18/06/2020

**Tóm tắt:**

*Bài báo này giới thiệu một phương pháp tóm tắt trích rút các văn bản sử dụng BERT. Để làm điều này, các tác giả biểu diễn bài toán tóm tắt trích rút dưới dạng phân lớp nhị phân mức câu. Các câu sẽ được biểu diễn dưới dạng vector đặc trưng sử dụng BERT, sau đó được phân lớp để chọn ra những câu quan trọng làm bản tóm tắt. Chúng tôi thử nghiệm phương pháp trên 3 tập dữ liệu với 2 ngôn ngữ (Tiếng Anh và Tiếng Việt). Kết quả thực nghiệm cho thấy phương pháp cho kết quả tốt so với các mô hình khác.*

***Từ khóa:*** *Tóm tắt văn bản, xử lý ngôn ngữ, học máy, học sâu, học không giám sát.*

**Chữ viết tắt**

| TT | Chữ viết tắt | Ý nghĩa |
|---|---|---|
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| L2R | Learning to rank | Học để xếp hạng |
| TF-IDF | Term Frequency - Inverse Document Frequency | Là một kỹ thuật khai phá dữ liệu văn bản |
| MLP | Multi-layer Perceptron | Perceptron nhiều lớp |

## 1. Giới thiệu

Tóm tắt văn bản tự động là một nhiệm vụ đầy thách thức nhưng thú vị của xử lý ngôn ngữ tự nhiên (NLP). Nhiệm vụ đặt ra là tạo ra một bản tóm tắt súc tích trong đó lưu trữ hầu hết thông tin từ một hoặc nhiều tài liệu. Công việc này được bắt đầu từ những năm 1950 [1]. Đầu ra của một hệ thống tóm tắt văn bản mang lại lợi ích cho nhiều ứng dụng NLP như tìm kiếm Web. Công cụ tìm kiếm Google thường trả về một đoạn mô tả ngắn về các trang Web tương ứng với truy vấn tìm kiếm, hoặc nhà cung cấp tin tức trực tuyến cung cấp các điểm nổi bật của tài liệu Web trên giao diện của nó. Điều này đòi hỏi các hệ thống tóm tắt văn bản chất lượng cao.

Về mặt kỹ thuật tóm tắt văn bản có hai hướng nghiên cứu: học có giám sát và học không giám sát. Hướng thứ nhất cần gán nhãn dữ liệu được huấn luyện bởi một bộ phân loại, điều này có thể quyết định xem một câu có nên được đưa vào bản tóm tắt hay không. Trong huấn luyện, các phương pháp học có giám sát sử dụng các đặc trưng được xác định trước bằng tay, trích xuất từ dữ liệu để huấn luyện mô hình sử dụng dự đoán các đầu vào chưa biết [2-5]. Cách tiếp cận này phù hợp với dữ liệu được gán nhãn đúng và có các đặc trưng phù hợp. Tuy nhiên, trên thực tế, dữ liệu được gán nhãn thường không có sẵn và việc xác định các đặc trưng phù hợp cho một miền cụ thể cũng là một nhiệm vụ đầy thách thức. Điều này gợi ý cho hướng nghiên cứu thứ hai với các phương pháp học không giám sát [6-11]. Phương pháp này khác với phương pháp học có giám sát ở chỗ chúng không cần huấn luyện dữ liệu và do đó dễ dàng thích ứng với các tên miền mới.

Những thành công gần đây của các mô hình biến đổi (transformers) mở ra hướng tiếp cận mới cho bài toán tóm tắt văn bản. Trong bài báo này chúng tôi giới thiệu một mô hình tóm tắt trích chọn câu dựa trên BERT (Bidirectional Encoder Representations from Transformers). Những đóng góp chính của bài báo này như sau:

- Bài báo đề xuất mô hình tóm tắt văn bản dựa trên BERT. Mô hình cho phép sử dụng sức mạnh của BERT được huấn luyện trên tập dữ liệu lớn, sau đó được áp dụng cho bài toán tóm tắt văn bản.
- Bài báo so sánh kết quả của mô hình với các phương pháp khác. Kết quả cho thấy mô hình tóm tắt dựa trên BERT cho kết quả khả quan.

Phần còn lại của bài viết này được tổ chức như sau. Phần II cung cấp các nghiên cứu liên quan. Phần III giới thiệu mô hình. Tiếp theo, Phần IV trình bày kết quả thực nghiệm và thảo luận. Cuối cùng, Phần VI đưa ra kết luận và hướng phát triển.

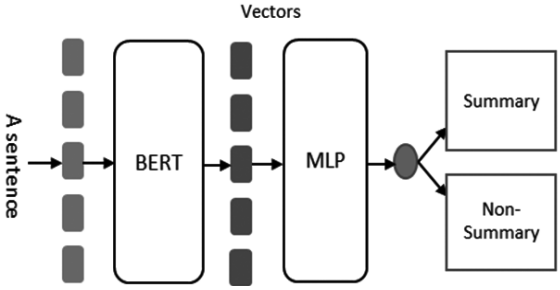
## 2. Các nghiên cứu liên quan

Các nghiên cứu tóm tắt văn bản đã được trình bày trong tài liệu [1], [2], [7], [15]. Các nghiên cứu đã có những cách tiếp cận bài toán tóm tắt văn bản với một số dạng khác nhau: bài toán xếp hạng (ranking problems) [1], bài toán tối ưu hóa [7], học có giám sát với bài toán phân loại nhị phân dựa trên tập đặc trưng ([2], [15]) hoặc phương pháp phân tích ma trận ([11], [16], [17]). Thành công gần đây của học sâu đã thu hút các nhà nghiên cứu trong việc áp dụng kỹ thuật này vào bài toán tóm tắt văn bản. Các kết quả đạt được nhờ tận dụng sự phân cấp [4], hoặc dựa trên đặc tính tuần tự của các câu [5].

Có một số nghiên cứu về tóm tắt văn bản bằng Tiếng Việt. Nguyễn-Hoàng và cộng sự đã giới thiệu một mô hình dựa trên đồ thị để trích xuất các câu quan trọng [12]. Các tác giả đã định nghĩa một phần mở rộng của TF-IDF để đo độ tương tự giữa hai câu phù hợp. Điểm này đã được sử dụng để tạo ra trọng lượng của các cạnh trong biểu đồ. Tóm tắt được trích xuất bằng cách xếp hạng câu với điểm số của nó. Tác giả và cộng sự đã nghiên cứu cụm từ chấm điểm trong vấn đề tóm tắt văn bản [14]. Các tác giả đề xuất sự kết hợp của các phương pháp tương tự khác nhau: TRComparer, TF-IDF, và Skip-Th Think vector. Văn bản tóm tắt được tạo ra bằng cách sử dụng một thuật toán xếp hạng. Nghiên cứu phù hợp nhất với nghiên cứu của chúng tôi là bài báo của tác giả Ung và cộng sự. [13]. Các tác giả đã trình bày một cách tiếp cận để tóm tắt đa tài liệu Tiếng Việt bằng cách kết hợp đặc trưng của câu. Độ quan trọng của một câu được tính bằng cách tính tổng trọng số của các đặc trưng. Các câu xếp hạng điểm cao của một chủ đề đã được chọn làm tóm tắt. Những phương pháp này đạt được kết quả đầy hứa hẹn, tuy nhiên, đây là phương pháp không giám sát. Khác với các nghiên cứu trước, trong bài báo này, chúng tôi đã tiến hành thực nghiệm và so sánh một số các phương pháp tóm tắt văn bản theo ba cách tiếp cận: không giám sát, giám sát và học sâu [12-14].

## 3. Mô hình

Chúng tôi giới thiệu mô hình tóm tắt trích rút dựa trên BERT ở Hình 1. Các từ trong một câu đầu vào sẽ được biến đổi bằng BERT để có được một vector đầu ra. Vector này là biểu diễn của câu đầu vào (low-dimensional vector). Vector này sẽ là đầu vào của một mạng truyền thẳng (feed-forward network). Mạng này sẽ cho ra một vector cuối cùng cho phân lớp. Kết quả ở bộ phân lớp cho biết câu đó có là câu tóm tắt hay không.




*Hình 1. Mô hình tóm tắt trích rút dựa trên BERT*

### 3.1. Mô hình biến đổi (Transformers)

Cấu trúc biến đổi (transformer) dựa trên "sự tự chú ý" (self attention) để tính toán sự biểu diễn của dữ liệu đầu vào mà không sử dụng cấu trúc mạng neural hồi quy (recurrent neural networks - RNN) [18]. Cấu trúc biến đổi dựa trên bộ mã hoá và giải mã (encoder - decoder) kết hợp với cơ chế "self attention", và các lớp mạng neural truyền thẳng. Cơ chế "attention" cho phép cấu trúc biến đổi tính toán sự ánh xạ của một truy vấn (query) và một tập các từ khoá - giá trị (key - value) cho đầu ra. Sau đó, đầu ra được tính toán bằng cộng có trọng số của các giá trị.

### 3.2. BERT

BERT là một mô hình dựa trên sự biến đổi (transformer), cho phép biểu diễn ngữ cảnh của một từ bằng cách dựa trên mối quan hệ của từ đó với các từ xung quanh [19]. BERT khác biệt với các mô hình một chiều (unidirectional) khi chỉ học các biểu diễn từ trái qua phải hoặc từ phải qua trái. BERT được sử dụng để huấn luyện mô hình ngôn ngữ mặt nạ (masked language model) bằng cách học hai bài toán cùng một lúc là bài toán dự đoán từ và bài toán dự đoán câu. Với bài toán dự đoán từ, các từ trong một câu sẽ được che giấu (masked). Quá trình huấn luyện sẽ dự đoán từ bị che dấu bằng cách dựa vào các từ xung quanh.

### 3.3. Phân lớp

Bộ phân lớp sử dụng vector đầu ra của MLP cho quá trình phân lớp. Chúng tôi sử dụng hàm *softmax()* cho quá trình phân lớp. Hàm này trả ra xác suất trên 2 tập nhãn (tóm tắt và không-tóm tắt).

### 3.4. Chọn câu tóm tắt

Các câu của một văn bản sau quá trình phân lớp sẽ được sắp xếp theo thứ tự giảm dần của độ quan trọng dựa trên xác suất dự đoán của mô hình. Thuật toán lựa chọn sẽ lấy m câu có thứ tự cao nhất làm bản tóm tắt.

### 3.5. Huấn luyện

Chúng tôi sử dụng mô hình BERT [19] cho quá trình huấn luyện. BERT sử dụng 12 tầng "attention" với 12 đầu (heads), và 177 triệu tham số. Chúng tôi sử dụng mô hình huấn luyện sẵn của BERT trên 102 ngôn ngữ, trong đó có Tiếng Việt. Quá trình huấn luyện trong 20 lần lặp với hệ số lỗi là 5x10-5 trên một GPU.

## 4. Kết quả thực nghiệm và thảo luận

### 4.1. Dữ liệu

Chúng tôi sử dụng 3 bộ dữ liệu để đánh giá mô hình, trong đó hai bộ Tiếng Anh và một bộ Tiếng Việt. **SoLSCSum** gồm 157 văn bản được thu thập từ Yahoo News [20] được sử dụng cho tóm tắt văn bản sử dụng ý kiến người dùng. Các câu trong văn bản được gán nhãn bằng tay. **USAToday-CNN** là bộ dữ liệu được thu thập từ hai trang USAToday và CNN, gồm 121 văn bản tương ứng với 121 sự kiện [24]. Các văn bản gồm hai phần: văn bản và các tweets liên quan. Các câu trong văn bản được gán nhãn bởi người gán nhãn. **VSoLSCSum** là bộ dữ liệu Tiếng Việt cho tóm tắt văn bản sử dụng ý kiến người dùng [21]. Chúng tôi sử dụng ba bộ dữ liệu này do các câu trong văn bản đã được gán nhãn, điều này thuận lợi cho quá trình huấn luyện mô hình tóm tắt.

### 4.2. Thiết đặt thực nghiệm

Chúng tôi sử dụng phương pháp *k*-fold cross-validation, tức là bộ dữ liệu sẽ được chia thành *k* phần bằng nhau, sau đó mô hình sẽ được lần lượt huấn luyện trên *k-1* phần và kiểm tra trên phần còn lại. Quá trình lặp lại *k* lần. Kết quả cuối cùng của mô hình là trung bình cộng điểm ROUGE trên toàn bộ các phần.

Với bộ dữ liệu **SoLSCSum**, chúng tôi sử dụng *k=10* [20], với bộ dữ liệu **USAToday-CNN**, chúng tôi sử dụng *k=5* [22], và với bộ dữ liệu VSoLSCSum chúng tôi sử dụng *k=5* [21]. Chúng tôi đặt *m=6* với hai bộ SoLSCSum và VSoLSCSum, và *m=4* cho bộ USAToday-CNN. Mô hình được huấn luyện trong 20 lần lặp (20 epochs) sử dụng hàm mất mát là cross-entropy.

### 4.3. Độ đo đánh giá

Chúng tôi sử dụng độ đo ROUGE[29] (Recall-Oriented Understudy for Gisting Evaluation) để so sánh kết quả của các mô hình. Các câu trích chọn sẽ được so khớp với các câu chuẩn (glod-data) trong tập dữ liệu để tính điểm ROUGE. ROUGE dựa vào sự giống nhau trên các từ (n-grams) để tính toán ra điểm. Vậy bản tóm tắt có điểm ROUGE càng cao thì càng tốt. Chúng tôi sử dụng pyrouge với tham số "-c 95 -2 -1 -U -r 1000 -n 2 -w 1.2 -a -s -f B –m".

$$ROUGE - N = \frac{\sum_{s \epsilon S_{ref}} \sum_{gram_n \epsilon s} Count_{match}(gram_n)}{\sum_{s \epsilon S_{ref}} \sum_{gram_n \epsilon s} Count(_{gram_n})}$$

Trong bài báo này, chúng tôi sử dụng ROUGE-1 và ROUGE-2 cho quá trình so sánh.

- ROUGE-1: tính toán sự giống nhau dựa trên các từ đơn (uni-gram).
- ROUGE-2: tính toán sự giống nhau trên 2 từ liên tục (bi-gram).

### 4.4. So sánh kết quả

Chúng tôi so sánh kết quả mô hình đề xuất với các mô hình khác trên ba bộ dữ liệu. Các mô hình gồm: **Lead-m** lấy m câu đầu tiên của văn bản làm bản tóm tắt [25]. LexRank sử dụng thuật toán **LexRank** để lấy các câu quan trọng [25]. **HGRW** là mô hình sử dụng thông tin từ ý kiến người dùng để nâng cao chất lượng bản tóm tắt [26]. **SoRTESum** là mô hình sử dụng các thuộc tính để xây dựng mô hình tính điểm dựa vào thông tin ý kiến người dùng [27]. **SVMRank** dựa trên ý tưởng của [23] để xây dựng mô hình tính điểm bằng các đặc trưng. **CNN** là mô hình tóm tắt dựa trên mạng tích chập (convolutional neural networks).

*Bảng 1. Kết quả so sánh trên ba bộ dữ liệu*
*(Chữ đậm thể hiện mô hình tốt nhất, chữ nghiêng thể hiện mô hình tốt thứ hai.)*

| Mô hình | SoLSCSum | | USAToday-CNN | | VSoLSCSum | |
|---|---|---|---|---|---|---|
| | ROUGE-1 | ROUGE-2 | ROUGE-1 | ROUGE-2 | ROUGE-1 | ROUGE-2 |
| Lead-*m* | 0.345 | *0.322* | 0.249 | **0.106** | 0.495 | 0.420 |
| LexRank | 0.327 | 0.243 | 0.251 | 0.092 | 0.506 | 0.432 |
| HGRW | 0.379 | 0.204 | **0.279** | *0.098* | 0.570 | 0.469 |
| SoRTESum | 0.352 | 0.277 | 0.252 | 0.085 | 0.532 | 0.463 |
| SVMRank | *0.401* | *0.322* | 0.253 | 0.084 | *0.576* | **0.523** |
| CNN | **0.413** | **0.367** | 0.214 | 0.071 | 0.543 | 0.447 |
| Our model | 0.360 | 0.290 | *0.270* | 0.081 | **0.602** | *0.478* |

Kết quả từ Bảng 1 cho thấy mô hình của chúng tôi cho kết quả khả quan trên ba bộ dữ liệu. Trên bộ VSoLSCSum, mô hình đạt kết quả tốt nhất với ROUGE-1, hơn mô hình thứ 2 (SVM Rank) khoảng 3%. Với ROUE-2, mô hình cho độ tốt thứ 2, sau SVM Rank. Bên cạnh đó, mô hình cũng đứng thứ 2 với ROUGE-1 trên bộ USAToday-CNN. Điều này là do mô hình tận dụng sức mạnh của BERT, được huấn luyện trên bộ dữ liệu rất lớn. Quá trình huấn luyện cho phép biểu diễn yếu tố ngữ cảnh của từ trong câu. Điều này cho phép mô hình chúng tôi có thể dự đoán đúng những câu quan trọng. Trên bộ SoLSCSum, mô hình không cho kết quả tốt nhất. Điều này có thể do mô hình chưa học được mẫu biểu diễn trên tập dữ liệu này. Kết quả từ Bảng 1 cũng cho thấy SVM Rank là một mô hình mạnh, cho kết quả ổn định trên cả 3 bộ dữ liệu [23]. Điều này là do mô hình tóm tắt dựa trên SVM Rank được huấn luyện với một tập các đặc trưng từ ba nguồn: đặc trưng từ câu, đặc trưng từ các ý kiến người dùng, và đặc trưng từ các văn bản liên quan thu thập từ Google. Mô hình dựa trên CNN cho kết quả tốt trên SoLSCSum nhưng không đạt kết quả tốt trên hai bộ còn lại. Điều này có thể do mô hình CNN phù hợp với bộ dữ liệu SoLSCSum. Hai mô hình tóm tắt không giám sát SoRTESum và HGRW cho kết quả khả quan trên cả ba bộ dữ liệu. Điều này cho thấy trong một số trường hợp, các mô hình học không giám sát có thể cho kết quả tương đương với các mô hình học có giám sát. Lead-m cho kết quả tốt nhất với ROUGE-2 trên bộ USAToday-CNN. Điều này cho thấy mặc dù mô hình đơn giản nhưng phản ánh đúng dữ liệu văn bản tin tức, trong đó thông tin quan trọng thường được viết ở những câu đầu tiên [24].

## 5. Kết luận

Bài báo này giới thiệu một mô hình tóm tắt trích rút câu cho đơn văn bản dựa trên BERT. Mô hình sử dụng BERT để tận dụng khía cảnh ngữ cảnh của các từ do BERT được huấn luyện trên một tập dữ liệu lớn. Bằng cách sử dụng BERT, mô hình có thể biểu diễn tốt mối quan hệ của các từ trong một câu, từ đó nâng cao chất lượng quá trình học các mẫu tiềm ẩn trong dữ liệu. Qua đó, giúp cải tiến quá trình phân lớp. Kết quả thực nghiệm trên ba bộ dữ liệu ở hai ngôn ngữ Tiếng Anh và Việt cho thấy mô hình cho kết quả khả quan so với các mô hình khác.

## 6. Cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Sư phạm Kỹ thuật Hưng Yên trong đề tài mã số UTEHY.L.2020.04.

## Tài liệu tham khảo

[1]. H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research Development*, **2(2)**, pp. 159-165, 1958.

[2]. D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *IJCAI*, pp. 2862-2867, 2007.

[3]. K. Hong and A. Nenkova, "Improving the estimation of word importance for news multi-document summarization," in *EACL*, pp. 712-721, 2014.
[4]. Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in *AAAI*, pp. 2153-2159, 2015.
[5]. P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, "Leveraging contextual sentence relations for extractive summarization using a neural a ention model," in *SIGIR*, 2017.
[6]. G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, **22**, pp. 457-479, 2004.
[7]. K. Woodsend and M. Lapata, "Automatic generation of story highlights," in ACL: 565-574, 2010.
[8]. J. A. B. Hui Lin, "A class of submodular functions for document summarization," in *ACL*, pp. 510-520, 2011, June.
[9]. K. Woodsend and M. Lapata, "Multiple aspect summarization using integer linear programming," in *EMNLP-CoNLL*, pp. 233-243, 2012.
[10]. S. Banerjee, P. Mitra, and K. Sugiyama, "Multi-document abstractive summarization using ilp based multi-sentence compression," in *IJCAI*, pp. 1208-1214, 2015.
[11]. M.-T. Nguyen, T. V. Cuong, N. X. Hoai, and M.-L. Nguyen, "Utilizing user posts to enrich web document summarization with matrix cofactorization," in *SoICT*, pp. 70-77, 2017.
[12]. T.-A. Nguyen-Hoang, K. Nguyen, and Q.-V. Tran, "Tsgvi: a graph-based summarization system for vietnamese documents," *Journal of Ambient Intelligence and Humanized Computing*, **3(4)**, pp. 305-312, 2012.
[13]. V.-G. Ung, A.-V. Luong, N.-T. Tran, and M.-Q. Nghiem, "Combination of features for vietnamese news multi-document summarization," in *The Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pp. 186-191, 2015.
[14]. H. Nguyen, T. Le, V.-T. Luong, M.-Q. Nghiem, and D. Dinh, "The combination of similarity measures for extractive summarization," in *Proceedings of the Seventh Symposium on Information and Communication Technology (SoICT)*, pp. 66-72, 2016.
[15]. J. Kupiec, J. O. Pedersen, and F. Chen, "A trainable document summarizer," in *SIGIR*, pp. 68-73, 1995.
[16]. D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *SIGIR*, pp. 307-314, 2008.
[17]. J.-H. Lee, S. Park, C.-M. Ahn, and D. Kim, "Automatic generic document summarization based on non-negative matrix factorization," Inf. *Process. Manage*, **45(1)**, pp. 20-34, 2009.
[18]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.
[19]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **Volume 1** (Long and Short Papers), pp. 4171-4186, 2019.
[20]. Minh-TienNguyen, Chien-XuanTran, Duc-VuTran, and Minh-LeNguyen, SoLSCSum: A Linked Sentence-Comment Dataset for Social Context Summariza- tion. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 2409-2412. ACM, 2016.
[21]. Minh-TienNguyen, VietDacLai, Phong-KhacDo, Duc-VuTran, and Minh-Le Nguyen, VSoLSCSum: Building a Vietnamese Sentence-Comment Dataset for Social Context Summarization. In *The 12th Workshop on Asian Language Resources*, pp. 38-48, 2016. Association for Computational Linguistics.

[22]. Minh-Tien Nguyen, Duc-Vu Tran, and Minh-Le Nguyen, Social Con- text Summarization using User-generated Content and Third-party Sources. *Knowledge-Based Systems*, **144**(2018), pp. 51-64. Elsevier, 2018.

[23]. Wei, Z. and Gao, W., Utilizing microblogs for automatic news highlights extraction. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pp. 872-883, 2014. Association for Computational Linguistics.

[24]. Ani Nenkova. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *AAAI*, **vol. 5**, pp. 1436-1441, 2005.

[25]. Gunes Erkan and Dragomir R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, **22**, pp. 457-479, 2004.

[26]. Zhongyu Wei and Wei Gao, Gibberish, Assistant, or Master?: Using Tweets Linking to News for Extractive Single-Document Summarization. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1003-1006. ACM, 2015.

[27]. Minh-Tien Nguyen and Minh-Le Nguyen, SoRTESum: A Social Context Framework for Single-Document Summarization. In *European Conference on Information Retrieval*, pp. 3-14. Springer International Publishing, 2016.

[28]. https://github.com/andersjo/pyrouge

## VIETNAMESE MULTI-DOCUMENT SUMMARIZATION BASE BERT METHODS

**Abstract:**

*This paper introduces an extractive summarization model based on BERT for news documents. To do that, we employ BERT to take into account the power of BERT trained on a huge amount of general data. This enables our model to effectively represent the hidden features of data. BERT is adapted to our summarization task by using transfer learning. The importance of sentences of an input document are estimated by using the trained model. A summary is formed by selecting top m ranked sentences based on margin probabilities. Experimental results on three datasets in two languages (English and Vietnamese) show that our model achieves promising results compared to strong methods.*

***Keywords:*** *text summarization, NLP, machine learning, deep learning, unsupervised learning.*