# XÁC ĐỊNH ĐẶC ĐIỂM TÁC GIẢ VĂN BẢN TIẾNG VIỆT BẰNG HỌC SÂU

**Dương Trần Đức**
*Học viện Công nghệ Bưu chính Viễn thông*

***Tóm tắt:*** Xác định đặc điểm tác giả văn bản là việc chỉ ra các đặc điểm của tác giả như giới tính, độ tuổi, .v.v chỉ dựa trên phân tích văn bản của tác giả đó. Bài báo này trình bày nghiên cứu về xác định đặc điểm tác giả văn bản tiếng Việt bằng phương pháp học sâu dựa trên mạng nơ ron tích chập (Convolutional Neural Network – CNN). Các thực nghiệm được thực hiện trên tập dữ liệu là các bài viết diễn đàn tiếng Việt đã được sử dụng trong các nghiên cứu trước đây về nhận diện đặc điểm tác giả văn bản tiếng Việt [8]. Kết quả thực nghiệm cho thấy phương pháp mới có kết quả nhận diện tốt hơn so với các phương pháp học máy truyền thống như Máy véc tơ hỗ trợ (Support Vector Machine) và Rừng ngẫu nhiên (Random Forest).

***Từ khóa:*** học sâu, mạng nơ ron tích chập, nhận diện đặc điểm tác giả.

## I. MỞ ĐẦU

Xác định đặc điểm tác giả văn bản (author profiling) là một nhánh nghiên cứu của phân tích tác giả văn bản. Phân tích tác giả văn bản còn có hai nhánh nghiên cứu khác là nhận diện tác giả (authorship attribution) và xác minh tác giả (author verification) [29]. Trong khi việc nhận diện tác giả hoặc xác minh tác giả tiến hành xác định hoặc kiểm chứng một tác giả cụ thể là người tạo nên văn bản và thường áp dụng cho các loại văn bản chính thống như bài báo, tiểu thuyết .v.v, xác định đặc điểm tác giả văn bản thường được thực hiện trên các loại văn bản tự do hơn như các loại văn bản trực tuyến (bài viết blog, email, diễn đàn .v.v) [1, 2, 5, 9, 12, 22, 29]. Do đó, các ứng dụng của xác định đặc điểm tác giả văn bản cũng khác so với hai nhánh nghiên cứu còn lại, vốn thường được sử dụng để giải quyết các tranh cãi về quyền tác giả. Ứng dụng chủ yếu của xác định đặc điểm tác giả là trong các lĩnh vực quảng cáo trực tuyến, cá nhân hóa hệ thống web, hỗ trợ điều tra tội phạm trực tuyến .v.v, trong đó các đặc điểm cá nhân của tác giả bài viết được dự đoán để hỗ trợ các hoạt động quảng cáo đúng mục đích hoặc điều tra tội phạm.

Cùng với sự phát triển của Internet và các kênh trao đổi thông tin trực tuyến, ứng dụng của việc xác định đặc điểm tác giả văn bản càng trở nên cần thiết và quan trọng hơn. Các nghiên cứu trước đây về xác định đặc điểm tác giả văn bản thường sử dụng các phương pháp học máy truyền thống trên tập các véc tơ đặc trưng. Một số phương pháp học máy truyền thống được sử dụng phổ biến cho các nghiên cứu xác định đặc điểm tác giả văn bản bao gồm SVM [2, 3, 5, 7, 13, 17, 21, 28], Logistic Regression [15, 16], Random Forest [29], Multi-Class Real Winnow [4]. Các tập đặc trưng được thử nghiệm được chia thành hai loại là đặc trưng dựa theo phong cách và dựa theo nội dung. Phong cách viết được xem như là một phương pháp độc lập miền và được sử dụng trong nhiều nghiên cứu trước đây về xác định đặc điểm tác giả. Hầu hết các thành phần có tính độc lập nội dung của ngôn ngữ đã được sử dụng làm đặc trưng phong cách như các ký tự, tính chất từ, từ loại [4, 5, 14, 27], từ công cụ (từ chức năng) [2, 6, 11], các cấu trúc ngữ pháp [5, 6, 23] .v.v. Các đặc trưng này thường được tạo ra từ các quy tắc của ngôn ngữ và không phụ thuộc vào tập dữ liệu hay lĩnh vực cụ thể nào. Ngược lại, các từ nội dung thường được lựa chọn từ chính các tập dữ liệu được sử dụng trong nghiên cứu hoặc được lựa chọn từ các từ ngữ có ngữ nghĩa liên quan đến lĩnh vực cụ thể [2, 10, 12]. Do đó, các từ nội dung được xem là có tính phụ thuộc miền hoặc phụ thuộc dữ liệu ở mức độ nào đó.

Các nghiên cứu trước đây thường được thực hiện trên các tập dữ liệu khác nhau (về ngôn ngữ, đặc điểm phân tích, độ lớn, .v.v). Do vậy, khó để xác định phương pháp được đề xuất nào là tốt nhất. Trong những năm gần đây, cuộc thi PAN về phân tích tác giả văn bản đóng vai trò quan trọng trong lĩnh vực nghiên cứu này. Việc các nghiên cứu trong cuộc thi được thực hiện và so sánh trong cùng điều kiện (tập dữ liệu, các tiêu chí) đem lại sự đánh giá khách quan cho các phương pháp và các xu hướng mới. Những năm đầu của cuộc thi PAN (2013-2016), các nghiên cứu tham gia hầu hết thực nghiệm trên các phương pháp học máy truyền thống và trên các tập đặc trưng đa dạng, trong đó SVM vẫn là phương pháp nổi bật và đem lại những kết quả tốt nhất. Trong cuộc thi năm 2017-2018, mặc dù các phương pháp học máy truyền thống như SVM vẫn tiếp tục được nhiều nghiên cứu lựa chọn, các phương pháp mới như học sâu bắt đầu được sử dụng và đem lại các kết quả tiềm năng [23, 24].

Đối với ngôn ngữ tiếng Việt, mặc dù đã có một số nghiên cứu về xác định đặc điểm tác giả văn bản trong ngôn ngữ này [6, 8], nhưng còn khá hạn chế. Ngoài ra, chưa có nghiên cứu nào về ứng dụng học sâu cho xác định đặc điểm tác giả văn bản tiếng Việt. Nghiên cứu này được thực hiện với mục tiêu thử nghiệm phương pháp học sâu cho xác định đặc điểm tác giả văn bản tiếng Việt (thực nghiệm trên tập dữ liệu bài viết diễn đàn tiếng Việt) và so sánh với các kết quả của các nghiên cứu trước đây về xác định đặc điểm tác giả văn bản tiếng Việt bằng các phương pháp học máy truyền thống.

Bài báo có cấu trúc như sau. Phần II trình bày về các nghiên cứu liên quan trong lĩnh vực phân tích tác giả và mạng nơ ron tích chập cho xử lý văn bản. Phần III mô tả phương pháp. Phần IV trình bày về các kết quả và thảo

Tác giả liên hệ: Dương Trần Đức,
Email: duongtranduc@gmail.com
Đến tòa soạn: 7/2019, chỉnh sửa: 8/2019, chấp nhận đăng: 8/2019.

luận. Cuối cùng, các kết luận sẽ được trình bày trong phần V của bài báo.

## II. TỔNG QUAN

### *A. Phân tích tác giả văn bản*

Phân tích tác giả văn bản là quá trình phân tích một tài liệu để có thể đưa ra các kết luận về tác giả của nó. Những nghiên cứu đầu tiên về phân tích tác giả xuất hiện từ đầu thế kỷ 19, với các phân tích về phong cách viết để nhận diện các tác phẩm của các tác giả như Shakespeare hay Bacon. Tuy nhiên, nghiên cứu được coi là chính thức đầu tiên trong lĩnh vực này được thực hiện bởi Mosteller và Wallace (1964) nhằm xác định tác giả của các bài luận cương liên bang (Federalist Papers) thông qua việc phân tích tần suất các từ chức năng được sử dụng trong văn bản. Nghiên cứu này khởi đầu cho một loạt các nghiên cứu tiếp theo về phân tích tác giả sử dụng các đặc trưng về "phong cách". Thời kỳ tiếp theo (từ cuối những năm 1990s), sự phát triển của Internet dẫn đến một số lượng lớn các văn bản trực tuyến được tạo ra, đồng thời các cải tiến về các mô hình tính toán như học máy đã thúc đẩy các nghiên cứu trong lĩnh vực này. Ngoài ra, các nghiên cứu cũng phát triển nhiều hơn theo nhánh xác định đặc điểm của các tác giả của các văn bản vô danh hơn là nhận diện tác giả của các văn bản chính thống.

Đối với các phương pháp phân tích truyền thống, quá trình phân tích tác giả văn bản liên quan đến hai vấn đề chính, đó là kỹ thuật phân tích và tập đặc trưng phân biệt. Các kỹ thuật phân tích trong thời kỳ đầu thường sử dụng các kỹ thuật khá đơn giản dựa trên thống kê [25] và ứng dụng chủ yếu trong việc hỗ trợ xử lý các tranh cãi về tác giả của các văn bản dài (bài báo, sách .v.v). Các nghiên cứu gần đây chủ yếu khai thác kỹ thuật học máy để tận dụng khả năng tính toán của máy tính. Rất nhiều các thuật toán học máy đã được nghiên cứu và thử nghiệm thành công cho việc phân tích tác giả như SVM, Decision Tree, Neural Networks .v.v. Tập đặc trưng có thể được xem như một phương pháp biểu diễn văn bản trên khía cạnh phong cách viết hoặc cách sử dụng từ. Theo Argamon et al. [2], có hai loại đặc trưng chính được sử dụng trong phân tích tác giả văn bản: đặc trưng về phong cách và đặc trưng dựa trên nội dung. Đặc trưng về phong cách bao gồm các đặc trưng liên quan đến ký tự, tính chất từ (lexical), cách sử dụng các cấu trúc ngữ pháp (syntactic), và các đặc trưng về cấu trúc văn bản. Đặc trưng dựa trên nội dung bao gồm các từ nội dung được sử dụng thường xuyên trong lĩnh vực đó hơn là các lĩnh vực khác. Các từ này thường được chọn theo phương pháp thống kê tần suất xuất hiện trong tập dữ liệu hoặc dựa trên ngữ nghĩa của từ. Các đặc trưng dựa trên các thành phần của hệ thống từ vựng đã được chứng minh là có hữu ích trong việc xác định đặc điểm tác giả văn bản trong nhiều nghiên cứu trước đây. Từ các thành phần cơ bản như các ký tự riêng lẻ [4, 5, 13, 27, 28], các cụm ký tự n-grams [3, 12, 15, 21], đến các đặc điểm của từ như loại từ, mức độ đa dạng của từ vựng [5, 6, 14, 25], các từ công cụ [2, 6, 10, 14, 16], và các từ nội dung [2, 9, 11, 19, 21, 29] đã được nghiên cứu sử dụng. Trong nghiên cứu đầu tiên được xem là hoàn chỉnh trong lĩnh vực này, Mosteller và Wallace (1964) sử dụng một số từ công cụ để giải quyết vấn đề tranh chấp trong việc xác định tác giả các bài luận liên bang (Federalist Papers). Sau đó, có rất nhiều các nghiên cứu tiếp theo trong lĩnh vực phân tích tác giả văn bản đã khai thác và xác minh tính hữu ích của các từ công cụ trong lĩnh vực này với số các từ được sử dụng từ 122 đến 645 từ. Các đặc trưng dựa trên ký tự và đặc điểm từ như các ký tự đơn lẻ/cụm ký tự, độ dài tự, loại từ, mức độ đa dạng trong dùng từ cũng được sử dụng phổ biến. De Vel et al. [7] sử dụng các đặc trưng như độ dài từ/câu, loại từ, tần suất các ký tự/loại ký tự, cùng với các đặc trưng ngữ pháp khác để phân biệt 156 emails trong tiếng Anh. Zheng et al. Abbasi và Chen [1] sử dụng 79 đặc trưng từ vựng trong tổng số 418 đặc trưng để phân tích tác giả các bài viết diễn đàn tiếng Anh và tiếng Ả rập. Các tác giả của sử dụng một tập đặc trưng hiệu quả dựa trên việc khai thác các đặc điểm về hình thái và chính tả tiếng Ả rập (chẳng hạn bổ sung thêm hai đặc trưng về phần kéo dài trong tiếng Arập). Iqbal et al. [11] sử dụng 419 đặc trưng bao gồm các đặc trưng dựa trên ký tự, đựa trên đặc điểm từ, đặc trưng ngữ pháp để xây dựng một loại "vân chữ viết" nhằm xác minh các tác giả email hỗ trợ điều tra tội phạm. Một số nghiên cứu cũng sử dụng các cụm kết hợp ký tự (n-grams) để làm đặc trưng phân loại. Stamatatos [25] nghiên cứu phương pháp sử dụng các cụm ký tự có độ dài biến đổi để giải quyết vấn đề nhận diện tác giả trên các bản tin Reuters của 50 tác giả khác nhau. Ý tưởng chính của phương pháp này là so sánh mỗi cụm ký tự với các cụm ký tự tương đồng và giữ lại các cụm ký tự nổi trội hơn. Peersman et al. [17] dự đoán tuổi và giới tính của người dùng chat dựa trên các đoạn chat thu thập từ mạng xã hội Netlog tại Bỉ. Tác giả sử dụng các cụm ký tự và từ làm đặc trưng phân loại. Các cụm 1 từ, 2 từ, 3 từ, 4 từ và các cụm 2 ký tự, 3 ký tự, 4 ký tự được trích từ tập dữ liệu và sau đó được chọn lọc bởi thuật toán lựa chọn đặc trưng khi-bình phương (chi-square).

Đối với thuật toán học sâu, việc ứng dụng trong phân tích văn bản nói chung và phân tích tác giả nói riêng đã được nghiên cứu nhiều hơn, điển hình là các công bố trong các cuộc thi PAN các năm 2016, 2017 [23, 24]. Khác với phương pháp học máy truyền thống, việc ứng dụng học sâu cho phân tích tác giả đòi hỏi việc chuyển đổi văn bản thành một ma trận số để có thể áp dụng quy trình huấn luyện trong mạng học sâu. Do vậy, các nghiên cứu về học sâu cho phân tích tác giả không khai thác các đặc trưng đa dạng như các phương pháp học máy truyền thống mà tìm cách chuyển đổi văn bản như đã nói ở trên, trong đó phổ biến nhất là việc sử dụng các tập nhúng từ hoặc tập nhúng ký tự. Vấn đề này sẽ được trình bày chi tiết hơn ở phần tiếp theo.

### *B. Mạng nơ ron tích chập cho xử lý văn bản*

Mạng nơ ron tích chập là một kỹ thuật đã được kiểm nghiệm và ứng dụng rộng rãi trong lĩnh vực nhận diện hình ảnh. Tuy nhiên, việc ứng dụng kỹ thuật này cho trong lĩnh vực xử lý văn bản trong thời gian đầu còn hạn chế. Vấn đề chính trong việc ứng dụng trực tiếp kỹ thuật này trong xử lý văn bản là việc biểu diễn nó thành dạng ma trận số tương tự như hình ảnh. Vấn đề này được giải quyết nhờ việc véc tơ hoá các từ và chia văn bản thành các đoạn có kích thước bằng nhau về số từ. Việc vec tơ hoá từ theo mô hình mã hoá one-hot (mã hoá kiểu 1-of-V, trong đó V là kích thước tập từ vựng) tỏ ra không hiệu quả do độ dài véc tơ quá lớn và không khai thác được mỗi liên quan ngữ nghĩa của các từ. Phương pháp véc tơ hoá từ được sử dụng phổ biến và hiệu quả hiện nay là tập nhúng từ (word embeddings). Phương pháp này sử dụng mạng nơ ron có 1 tầng ẩn với đầu vào là một tập dữ liệu lớn và

sinh ra một không gian véc tơ với số chiều nhỏ hơn rất nhiều so với kích thước tập từ vựng (chỉ khoảng vài trăm). Mỗi từ trong tập dữ liệu sẽ được gắn với 1 véc tơ trong không gian và các từ có cùng ngữ cảnh sẽ được đặt gần nhau trong không gian véc tơ [12]. Ngoài ra, khi chia văn bản thành các đoạn có kích thước bằng nhau, một số đoạn không có đủ kích thước có thể được đệm vào một số từ trống để cho đủ kích thước quy định. Các kỹ thuật xử lý này giúp cho đoạn văn bản có thể được biểu diễn bằng một ma trận số giống như các hình ảnh và có thể áp dụng phương pháp CNN trên các dữ liệu này. Một số nghiên cứu điển hình về phân tích văn bản sử dụng kỹ thuật này là các nghiên cứu [12, 23], trong đó các tác giả đã sử dụng mạng nơ ron tích chập để nhận diện đặc điểm giới tính và ngôn ngữ của tác giả của các bài viết mạng xã hội Twitter và cho kết quả khả quan về khả năng ứng dụng CNN trong xử lý văn bản.

Bên cạnh việc sử dụng tập nhúng từ, các tập nhúng ký tự cũng được ứng dụng khá rộng rãi, điển hình là nghiên cứu [24], trong đó tác giả nghiên cứu sử dụng mạng nơ ron tích chập trên các cụm ký tự để nhận diện đặc điểm tác giả của các bài viết ngắn.

## III. PHƯƠNG PHÁP

Tương tự một số nghiên cứu trước đây về ứng dụng CNN cho xử lý văn bản, nghiên cứu này áp dụng kỹ thuật tập nhúng từ để tạo véc tơ từ và tiến hành chia văn bản thành các đoạn đều nhau (đệm từ trống cho các đoạn thiếu từ).

Cụ thể, các văn bản sẽ được thực hiện tách từ bằng công cụ tách từ có sẵn [18], sau đó chia thành các đoạn có kích thước $k$ từ. Các từ sau đó được biểu diễn bằng một véc tơ có độ dài $e$ theo kỹ thuật word embedding. Khi đó, mỗi đoạn văn bản sẽ được biểu diễn bởi một ma trận $C \in \mathbb{R}^{e \times k}$, trong đó mỗi cột tương ứng với một véc tơ từ. Ma trận này có thể được sử dụng làm đầu vào cho một CNN. Mạng này sẽ áp dụng các bộ lọc tích chập (convolutional filters) là các cụm từ với số lượng từ khác nhau. Giả sử một bộ lọc $H \in \mathbb{R}^{e \times w}$ được áp dụng trên một phần của $C$ (từ từ thứ $i$ đến từ thứ $w$, ký hiệu $C[i : i + w -1]$), với $w$ là kích thước bộ lọc (cũng là số từ của cụm từ). Ma trận kết quả $O$ được sử dụng làm đầu vào cho hàm sigmoid $g$, cùng với số bias $b$ để tạo ra đặt trưng $f_i$ của văn bản [24].

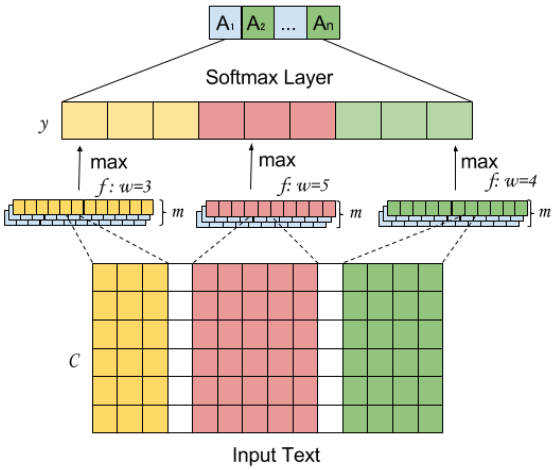
$$O = H \,.\, C[i : i + w -1] \quad (1)$$

$$f_i = g\,(O + b) \quad (2)$$

Bộ lọc này áp dụng trên các phần khác nhau có thể của $C$ sẽ tạo ra một bản đồ đặc trưng (feature map)

$$f = [f_1, f_2, ..., f_{k-w+1}] \quad (3)$$

Hình 1 cho thấy cấu trúc một CNN với số bộ lọc $m$ = 500 và kích thước bộ lọc lần lượt là $w$ = 2, 3, và 4 từ. Lưu ý rằng số hàng của bộ lọc và của ma trận đầu vào luôn luôn bằng nhau và là kích thước của véc tơ từ. Tiếp theo, thao tác max-pooling over time sẽ được thực hiện trên các bản đồ đặc trưng đầu ra. Theo đó, chỉ có giá trị lớn nhất của mỗi bản đồ đặc trưng $\hat{f} = \max\{f\}$ được sử dụng để làm đặc trưng tương ứng với bộ lọc đó. Ý tưởng của việc này là lấy giá trị quan trọng nhất (giá trị lớn nhất) của mỗi bản đồ đặc trưng. Việc chỉ lấy giá trị lớn nhất cũng cho phép quá trình có thể thực hiện trên nhiều bộ lọc kích thước khác nhau (số từ khác nhau).




*Hình 1. Mạng CNN cụm từ với các lớp lọc tích chập, max-pooling, và softmax [23].*

Như vậy, mỗi bộ lọc sẽ tạo ra một đặc trưng. Các đặc trưng này sẽ kết hợp với nhau thành một véc tơ và cuối cùng lớp kết nối đầy đủ (fully connected) softmax sẽ được sử dụng để thực hiện dự đoán đầu ra của CNN.

Để tối ưu hoá kết quả của CNN này, ba tham số sẽ được tuỳ chỉnh. Đầu tiên là kích thước của các đoạn văn bản đầu vào. Kích thước đầu vào (tính theo số từ) nhỏ quá sẽ khó đạt hiệu quả, trong khi kích thước lớn quá làm giảm số mẫu và làm tăng độ phức tạp thực hiện. Tham số thứ hai là số bộ lọc $m$ và kích thước bộ lọc $w$. Số bộ lọc lớn sẽ tăng khả năng đại diện, tuy nhiên dễ dẫn đến trình trạng quá khớp. Kích thước bộ lọc lớn có thể giúp nắm bắt mối quan hệ rộng giữa các từ, với điều kiện kích thước tập dữ liệu phải lớn tương ứng [24].

Ngoài ra, các tập nhúng từ sẽ được thực nghiệm theo hai loại là tập nhúng từ tạo trước (pre-trained) và tập nhúng từ được tạo trong quá trình huấn luyện mô hình. Tập nhúng từ tạo trước được sử dụng là tập các véc tơ từ đã được huấn luyện trước đó trên bộ dữ liệu tiếng Việt thu thập từ trang Wikipedia tiếng Việt. Tập nhúng từ tạo trong quá trình huấn luyện mô hình là các véc tơ từ được huấn luyện dựa trên tập dữ liệu được sử dụng trong nghiên cứu.

## IV. THỰC NGHIỆM VÀ KẾT QUẢ

### A. Dữ liệu

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu của nghiên cứu trước đây về nhận diện đặc điểm tác giả bài viết diễn đàn [8] để tiện so sánh kết quả. Tập dữ liệu này được thu thập bằng cách sử dụng bộ thu thập dữ liệu tự động (crawler) để thu thập các bài viết từ các diễn đàn phổ biến ở Việt Nam như otofun.net.vn, webtretho.com, tinhte.vn. Do các bài viết diễn đàn được viết khá tự do và chứa nhiều nội dung nhiễu, các phương pháp lọc và làm sạch dữ liệu đã được thực hiện như đã nói ở trên. Sau bước xử lý và làm sạch, tập dữ liệu thu thập được bao gồm có 6.831 bài viết từ 104 người dùng. Tổng cộng có 736.252 từ và trung bình 107 từ/bài. Các bài viết được lựa chọn là các bài có ít nhất một thông tin về đặc điểm người viết, có thể dùng làm dữ liệu huấn luyện cho hệ thống. Độ dài của các bài viết cũng được giới hạn trong khoảng từ 250 đến 1.500 ký tự để loại bỏ các bài viết quá ngắn hoặc quá dài (bài viết quá dài có thể chứa các đoạn văn bản sao chép từ các nguồn khác). Bảng 1 cho thấy các thông số thông kê về tập ngữ liệu huấn luyện.

*Bảng 1. Thống kê về tập ngữ liệu huấn luyện*

| Đặc điểm tác giả | Số bài viết | Lớp đặc điểm | Tỷ lệ |
|---|---|---|---|
| Giới tính | 4.474 | Nam | 54% |
| | | Nữ | 46% |
| Độ tuổi | 3.017 | Ít hơn 22 | 21% |
| | | Từ 24 đến 27 | 27% |
| | | Nhiều hơn 32 | 52% |
| Vùng miền | 3.960 | Bắc | 57% |
| | | Nam | 43% |
| Nghề nghiệp | 3.453 | Kinh doanh, bán hàng | 36% |
| | | Kỹ thuật, công nghệ | 31% |
| | | Giáo dục, y tế | 33% |

*B. Kết quả và đánh giá*

Các thực nghiệm được thực hiện sử dụng thư viện Tensorflow. Với kỹ thuật Gradient Descent, độ chính xác được tính toán và so sánh trong 200 vòng (epochs) và mô hình có độ chính xác tốt nhất được lưu lại để làm kết quả thực nghiệm.

Các tham số được thử nghiệm để tối ưu trong các khoảng như sau:

- Kích thước đầu vào (tính theo số từ): Nhỏ nhất từ 32 cho đến lớn nhất 256 (mỗi lần tăng gấp đôi).
- Kích thước bộ lọc *w*: Các bộ lọc được thử nghiệm gồm {1, 2, 3}, {2, 3, 4}, và {3, 4, 5}.
- Số bộ lọc *m*: từ 300 đến 1.500. Lần lượt mỗi bộ lọc là 100, 200, 300, 400, và 500,

Ngoài ra, hai tham số khác cũng được áp dụng để tránh vấn đề quá khớp là tham số drop-out =0,5 và L2 regularization = 0,7. Tập dữ liệu được chia thành hai tập huấn luyện và kiểm tra với tỷ lệ tập kiểm tra là 10%.

*Bảng 2. Giá trị tham số tối ưu*

| Tham số | Giá trị tối ưu |
|---|---|
| Kích thước đầu vào | 128 |
| Kích thước bộ lọc | {3, 4, 5} |
| Số bộ lọc | 1000 |
| Số drop-out | 0,5 |
| L2 regularization | 0,7 |

Các tham số trên được thực hiện tối ưu qua các thực nghiệm và bảng 2 cho thấy kết hợp tốt nhất của các tham số.

Bảng 3 cho thấy kết qủa nhận diện trên tập kiểm tra với bộ tham số tốt nhất trong các trường hợp:

- Mạng nơ rơn tích chập từ với tập nhúng từ tạo trước (**WCNN Pre-trained**).
- Mạng nơ ron tích chập từ với tập nhúng từ tự huấn luyện (**WCNN Self-trained**).
- Thuật toán học máy truyền thống **SVM** trên tập đặc trưng đầy đủ (kết quả lấy từ nghiên cứu trước trên cùng tập dữ liệu [8])

*Bảng 3. Kết quả thực nghiệm*

| Đặc điểm tác giả | WCNN Pre-trained | WCNN Self-trained | SVM Full |
|---|---|---|---|
| Giới tính | 92.17 | 93.96 | 91.72 |
| Độ tuổi | 72.36 | 72.95 | 71.26 |
| Vùng miền | 84.34 | 84.85 | 84.28 |
| Nghề nghiệp | 62.07 | 62.23 | 61.43 |

Kết quả ở bảng 3 cho thấy phương pháp mới có kết quả tốt hơn phương pháp học máy truyền thống có kết quả tốt nhất của nghiên cứu trước là SVM khi thực nghiệm trên cùng tập dữ liệu. Kết quả này cho thấy tiềm năng của phương pháp, đặc biệt là khả năng tối ưu hơn nữa do hệ thống các tham số của phương pháp đa dạng và có nhiều ảnh hưởng đến kết quả nhận diện.

Đối với việc sử dụng các tập nhúng từ khác nhau, tập nhúng từ tự huấn luyện cho kết quả tốt hơn tập nhúng từ được huấn luyện từ trước. Điều này có thể được giải thích do tập nhúng từ tự huấn luyện sẽ có khả năng phản ánh sát thực mối quan hệ giữa các từ trong tập dữ liệu hiện tại hơn. Mặc dù vậy, kết quả trên tập nhúng từ được huấn luyện trước cũng vẫn có độ chính xác cao hơn phương pháp học máy truyền thống. Ngoài ra, việc sử dụng tập nhúng từ huấn luyện trước còn được xem là phương pháp độc lập dữ liệu hơn khi các véc tơ từ được tạo một cách độc lập với tập dữ liệu được dùng trong nghiên cứu.

## V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã trình bày phương pháp sử dụng học sâu dựa trên mạng nơ ron tích chập để xác định đặc điểm tác giả văn bản tiếng Việt. Đây là phương pháp có nhiều sự khác biệt so với các phương pháp học máy truyền thống khi ứng dụng trong xử lý văn bản và chưa có nhiều nghiên cứu trong lĩnh vực nhận diện đặc điểm tác giả văn bản, đặc biệt là đối với văn bản tiếng Việt.

Các kết quả thực nghiệm cho thấy độ chính xác nhận diện khi sử dụng phương pháp này tốt hơn so với phương pháp học máy truyền thống đã được thực nghiệm cho kết quả tốt nhất trên cùng tập dữ liệu là SVM.

Hướng phát triển tiếp theo có thể là tiến hành các nghiên cứu trên loại đầu vào khác như tập nhúng ký tự hoặc tập nhúng các cặp ghép n ký tự (n-grams) thay vì tập nhúng từ. Các loại đầu vào này cũng đã được thử nghiệm trên các ngôn ngữ khác và cho kết quả khả quan, nhưng chưa được thử nghiệm trên ngôn ngữ tiếng Việt. Ngoài ra, các tham số của mạng nơ ron tích chập cũng cần được bổ sung và mở rộng khoảng khi thực hiện tối ưu nhằm tìm ra bộ tham số tốt nhất. Do việc huấn luyện trên mạng nơ ron tích chập là một hoạt động tiêu tốn tài nguyên và thời gian, nghiên cứu này chưa thực hiện tối ưu một cách triệt để các tham số của thuật toán.

## TÀI LIỆU THAM KHẢO

[1] A. Abbasi, H. Chen, Applying authorship analysis to extremist-group Web forum messages, IEEE Intelligent Systems (2005)

[2] A. Abbasi, H. Chen, Writeprints: A Style-based approach to identity-level identification and similarity detection in cyberspace, ACM Transactions on Information Systems, 26 (2), pp: 1-29 (2008)

[3] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August (2003)

[4] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler, Automatically Profiling the Author of an Anonymous Text, Communications of the ACM , in press (2008)

[5] M. Corney, O. DeVel, A. Anderson, and G. Mohay, Gender-preferential text mining of e-mail discourse, In ACSAC'02: Proc. of the 18th Annual Computer Security Applications Conference, Washington, DC, pp : 21-27. (2002)

[6] P. Dang, T. Giang, and P. Son, Author profiling for Vietnamese blogs, International Conference on Asian Language Processing (2009)

[7] O. De Vel, A. Anderson, M. Corney, and G. Mohay, Mining e-mail content for author identification forensics, SIGMOD Record 30(4), pp. 55-64 (2001)

[8] D. Đức, P. Sơn, và T. Hạnh, Xác định đặc điểm tác giả bài viết diễn đàn tiếng Việt dựa trên âm tiết và vần, Chuyên san các công trình nghiên cứu, phát triển, và ứng dụng Công nghệ thông tin và Truyền thông, Bộ Thông tin và Truyền thông, số 17(37) (2017).

[9] S. Goswami, S. Sarkar, and M. Rustagi, Style-based analysis of bloggers' age and gender, In Proceedings of the Third International ICWSM Conference. The AAAI Press (2009)

[10] G. Gressel, P. Hrudya, K. Surendran, S. Thara, A. Aravind, and P. Prabaharan, Ensemble learning approach for author profiling, Notebook for PAN at CLEF (2014)

[11] F. Iqbal, Messaging Forensic Framework for Cybercrime Investigation. A Thesis in the Department of Computer Science and Software Engineering - Concordia University Montréal, Canada (2010)

[12] Y. Kim, Convolutional neural networks for sentence classification, In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, Association for Computational Linguistics, Doha, Qatar (2014)

[13] M. Koppel, S. Argamon, and A. R. Shimoni, Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), pp : 401-412 (2002)

[14] T. Kucukyilmaz, C. Aykanat, B. B. Cambazoglu, and F. Can, Chat mining: predicting user and message attributes in computer-mediated communication. Information Processing and Management, 44(4), pp - 1448-1466 (2008)

[15] D. Nguyen, Noah A. Smith, and Carolyn P. Rosé, Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics (2011)

[16] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (2013)

[17] C. Peersman, W. Daelemans, and L. V. Vaerenbergh, Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, pages 37–44, New York, NY, USA, 2011. ACM (2007)

[18] L. H. Phuong, N. T. M. Huyen, R. Azim, T. H. Vinh, A hybrid approach to word segmentation of Vietnamese texts, Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, (2008).

[19] F. Rangel, P. Rosso, M. Potthast, B. Stein, Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings. CLEF and CEUR-WS.org (2017).

[20] F. Rangel, and P. Rosso, Use of language and author profiling: Identification of gender and age. In Natural Language Processing and Cognitive Science, p. 177 (2013)

[21] J. Savoy, Authorship attribution based on specific vocabulary. ACM Trans. Inf. Syst. 30, 2 (2012)

[22] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, Effects of Age and Gender on Blogging. In 43 proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006)

[23] S. Sierra, M. Montes-y-Gómez, T. Solorio, and F. A. González, Convolutional Neural Networks for Author Profling. In Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings), Vol. 1866. CLEF and CEUR-WS.org (2017).

[24] P. Shrestha, S. Sierra, F. Gonzalez, M. Montes, P. Rosso, T. Solorio, Convolutional neural networks for authorship attribution of short texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 669–674. Association for Computational Linguistics, Valencia, Spain (2017)

[25] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic text categorization in terms of genre and author, Computational Linguistics 26(4), pp. 471-495 (2000)

[26] C. Zhang, and P. Zhang, Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA (2010)

[27] X. Zhang, J. Zhao, Y. Le Cun, Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems. pp. 649–657 (2015)

[28] R. Zheng, H. Chen, Z. Huang, and Y. Qin, Authorship Analysis in Cybercrime Investigation (Eds.): ISI 2003, LNCS 2665, pp : 59-73 (2003)

[29] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378–393 (2006)

## AUTHOR PROFILING FOR VIETNAMESE TEXT USING DEEP LEANING

***Abstracrt:*** Author profiling is the task of identify the characteristics of authors such as gender, age, etc. only based on analysis of their texts. This paper show reseach on author profiling of Vietnamese texts using deep learning based on Convolutional Neural Networks (CNN). The experiments were conducted on the datasets which was experimented in the previous research on author profiling of Vietnamese texts [8]. The experiments show that the new method has better results than the traditional machine learning methods such as SVM (Support Vector Machine) and Random Forest on author profiling task.

***Keywords:*** deep learning, convolutional neural network, author profiling.

**Dương Trần Đức** Tốt nghiệp Đại học KHTN, Đại học Quốc gia Hà Nội ngành Công nghệ thông tin năm 1999, Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004, và Tiến sỹ chuyên ngành Kỹ thuật máy tính tại Học viện Công nghệ Bưu chính Viễn thông năm 2018. Hiện đang công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông.

# CÁC PHƯƠNG PHÁP QUẢN LÝ NHIỄU TRONG TRUYỀN THÔNG D2D

**Nguyễn Thị Yến*, Đinh Thị Thái Mai**, Lê Nhật Thăng***

**Học viện Công nghệ Bưu chính Viễn Thông*
***Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội*

***Tóm tắt:*** Với sự gia tăng nhanh chóng về số lượng thiết bị cầm tay (đặc biệt là điện thoại thông minh), mạng di động truyền thống dần không thể đáp ứng được nhu cầu về dung lượng tốc độ ngày càng cao hay độ trễ yêu cầu ngày càng thấp. Trong bối cảnh này, truyền thông giữa thiết bị với thiết bị (D2D) được xem là một công nghệ hiệu quả trong việc tăng hiệu quả phổ và giảm tải bằng cách giảm lưu lượng dữ liệu di động trong mạng di động. Tuy nhiên, để đạt được nhiều lợi ích, truyền thông D2D phải sử dụng nguồn tài nguyên một cách linh hoạt. Điều này dẫn đến nhiễu giữa truyền thông D2D và truyền thông di động. Trong bài báo này, chúng tôi thực hiện phân tích, đánh giá hai phương pháp quản lý nhiễu: sử dụng vùng hạn chế nhiễu và sử dụng vùng ngăn chặn nhiễu giữa người dùng D2D và người dùng di động áp dụng cho đường xuống dưới kịch bản mạng di động tái sử dụng tần số một phần (Partial Frequency Reuse - PFR) trên kênh pha-đinh Rayleigh. Kết quả mô phỏng bằng công cụ Matlab cho thấy tính hiệu quả của từng phương pháp quản lý nhiễu qua việc cải thiện được dung lượng hệ thống khi so sánh với phương pháp thông thường.

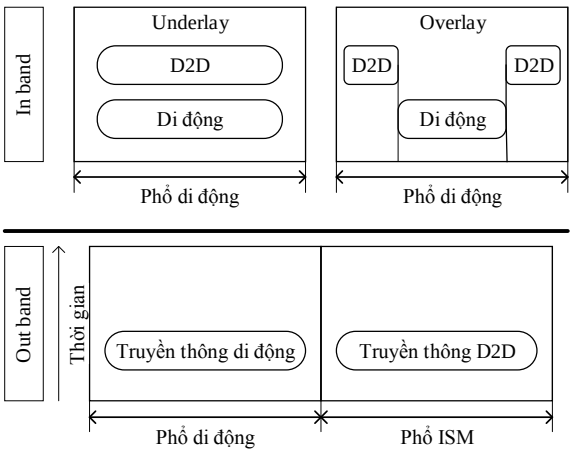
***Từ khóa:*** Mạng truyền thông D2D, quản lý nhiễu, phân bổ tài nguyên, U-D2D, SINR.

## I. GIỚI THIỆU

Trong thập kỷ qua, lưu lượng dữ liệu di động đã tăng lên đáng kể. Dự báo trong một vài năm tới đây, sự gia tăng này sẽ tiếp tục và nhiều gấp nhiều lần hơn nữa [1], điều này cho thấy tải trong mạng di động với kiến trúc truyền thống sẽ tăng lên và dần không đáp ứng được nhu cầu đặt ra. Để đáp ứng tải lưu lượng ngày càng tăng, truyền thông giữa thiết bị với thiết bị (D2D) [2], [3] đã được đề xuất. Truyền thông D2D ngày càng thu hút được sự quan tâm từ giới học thuật tới các ngành công nghiệp lớn nhằm giải quyết một loạt các vấn đề cấp bách mà mạng di động thông thường đang gặp phải như quá tải vì sự gia tăng nhanh chóng của các thiết bị di động hay không còn phù hợp với một số đòi hỏi về độ trễ của các dịch vụ mới.

Nhiều nghiên cứu đã chứng minh sự quan trọng của truyền thông D2D trong các mạng thế hệ tiếp theo (NGNs) [4], [5]. Các kết quả dựa trên phân tích và mô phỏng của các nghiên cứu này cho thấy những lợi ích vượt trội cho các ứng dụng như giảm tải và trễ cho mạng tế bào, tăng dung lượng kênh hay mở rộng vùng phủ sóng [6], [7]...

Về cơ bản, truyền thông D2D được chia thành hai hướng chính là truyền thông D2D sử dụng chung dải tần số với truyền thông di động (In band) và truyền thông D2D sử dụng khác dải tần số với truyền thông di động (Out band). Trong đó, truyền thông D2D Inband được chia thành 2 loại là Underlay (U-D2D) và Overlay (O-D2D). Hình 1 miêu tả sự khác biệt giữa hai phương pháp truyền thông D2D.




*Hình 1. Hai phương pháp truyền thông D2D*

Để có thể đạt được hiệu suất về dung lượng kênh, phương pháp dựa trên tái sử dụng tần số được xem là có hiệu quả nhất. Truyền thông D2D Inband-Underlay không phải là một ngoại lệ. Nguồn tài nguyên, cụ thể là các kênh tần số được tận dụng tối đa để cấp phát cho truyền thông D2D. Trong thực tế, trường hợp cặp liên kết D2D dùng chung tài nguyên với người dùng mạng di động (CUE) sẽ gây ra nhiễu [8]. Trong những năm vừa qua, nhiều thuật toán đã được đề xuất để giải quyết vấn đề này. Các phương pháp chủ yếu được sử dụng là điều khiển công suất và dựa trên chất lượng kênh truyền [9], [10]. Mục đích cuối cùng là làm sao tối đa được thông lượng của hệ thống mà vẫn đảm bảo được mức SINR cho người dùng di động. Một số nhóm tác giả đề xuất một giao thức mới như được đề cập ở [11].

Trong bài báo này, chúng tôi nghiên cứu và đánh giá hai phương pháp quản lý nhiễu trong truyền thông D2D khi xem xét đến ảnh hưởng của môi trường truyền lan trong không gian tự do: phương pháp quản lý nhiễu sử dụng vùng hạn chế nhiễu (ILA) và phương pháp quản lý nhiễu sử dụng vùng ngăn chặn nhiễu (ISA). Hai phương pháp này đều có chung ý tưởng xây dựng các vùng hạn

Tác giả liên hệ: Nguyễn Thị Yến,
Email: nguyenthiyen.nty281182@gmail.com
Đến tòa soạn: 7/2019, chỉnh sửa: 8/2019/2019, chấp nhận đăng: 8/2019.

chế nhiễu cho thiết bị D2D đầu cuối. Điểm khác biệt của hai phương pháp này chính là mô hình tính toán bán kính của vùng quản lý nhiễu. Sau đó, những nguồn tài nguyên sử dụng cho người dùng di động nằm trong bán kính vùng này sẽ được loại bỏ ra khỏi danh sách có thể cấp phát cho truyền thông D2D. Cuối cùng, hiệu năng hệ thống được đánh giá dưới tác động của kênh pha-đinh Reyleigh.

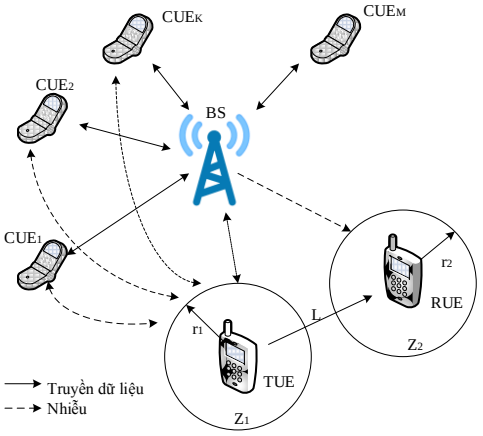
Phần còn lại của bài báo được tổ chức như sau: trong phần II, chúng tôi miêu tả mô hình, hoạt động của hệ thống truyền thông D2D. Trong phần III, IV chúng tôi trình bày cụ thể về các phương pháp quản lý nhiễu sử dụng vùng hạn chế nhiễu (ILA) và phương pháp quản lý nhiễu sử dụng vùng ngăn chặn nhiễu (ISA). Phần V giới thiệu về các kết quả mô phỏng và phân tích đánh giá, so sánh hiệu năng của các phương pháp quản lý nhiễu. Cuối cùng, kết luận bài báo sẽ được trình bày trong phần VI.

## II. MÔ HÌNH HỆ THỐNG TRUYỀN THÔNG D2D

Trong phần này, chúng tôi sẽ giới thiệu ngắn gọn về mô hình hệ thống của truyền thông D2D và cơ sở lý thuyết của các phương pháp quản lý nhiễu được nghiên cứu.

### *A. Mô hình hệ thống truyền thông D2D*

Chúng ta xét mô hình mạng gồm *M* người dùng di động (CUE) và một cặp truyền thông D2D. Chúng được phân bố một cách ngẫu nhiên trong tế bào và chịu sự quản lý của BS. Như có thể thấy trong Hình 2, người truyền D2D (TUE) truyền dữ liệu ở mức năng lượng $P_d$ tới người nhận D2D (RUE). Khoảng cách từ BS đến TUE và RUE lần lượt là $d_1$ và $d_2$. Khoảng cách giữa TUE và RUE là $L$. TUE được đặt trong vùng phủ $Z_1$ có bán kính $r_1$, RUE được đặt trong vùng phủ $Z_2$ có bán kính $r_2$.

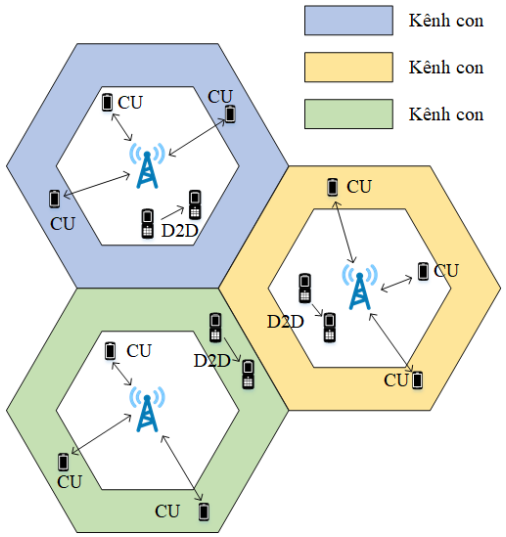



*Hình 2. Mô hình hệ thống truyền thông D2D*

Chúng tôi giả sử chỉ có *K* trong tổng số *M* CUE chia sẻ tài nguyên cho truyền thông D2D. Tín hiệu từ BS lúc này sẽ gây nhiễu cho cặp truyền thông D2D. Đồng thời, *K* CUE bị ảnh hưởng nhiễu từ cặp truyền thông D2D. Vì vậy, việc BS quản lý nhiễu giữa truyền thông D2D và mạng di động là rất cần thiết. Quy trình để hạn chế nhiễu trong phương pháp này được mô tả như sau. Đầu tiên, BS hạn chế nhiễu giữa truyền thông D2D và mạng di động bằng cách sử dụng phương pháp vùng hạn chế nhiễu. Sẽ không có CUEs nào sử dụng cùng tài nguyên với người dùng D2D trong các khu vực $Z_1$ và $Z_2$. Cuối cùng, BS quyết định các nguồn tài nguyên thích hợp cho người dùng D2D, nhằm cải thiện thông lượng mạng.

### *B. Hoạt động của hệ thống truyền thông D2D*

Truyền thông D2D được mô phỏng dưới kịch bản mạng tế bào gồm 3 trạm như Hình 3. Mạng di động sử dụng OFDMA kết hợp với công nghệ tái sử dụng tần số một phần (PFR) [13]. PFR được nghiên cứu trong mạng dựa trên OFDMA để khắc phục các vấn đề nhiễu đồng kênh. Trong PFR, vùng phủ của trạm gốc được phân chia thành vùng trung tâm và vùng biên, các tế bào sử dụng chung tần số cho vùng trung tâm và sử dụng các tần số đôi một khác nhau cho vùng biên và khác với vùng trung tâm. Trong mỗi vùng phủ của một trạm, người dùng ở trung tâm có thể sử dụng các kênh con trung tâm và biên, trong khi người dùng biên chỉ có thể sử dụng các kênh con ứng với vùng biên. Do đó, sự can thiệp giữa các tế bào đối với người dùng di động và người dùng D2D có thể gần như được loại bỏ và thông lượng hệ thống được cải thiện.




*Hình 3. Hoạt động của hệ thống truyền thông D2D*

Chúng tôi xác định thông lượng bằng cách áp dụng công thức Shannon [12]. Đối với mạng di động có chứa cặp truyền thông D2D, dung lượng mạng bằng tổng dung lượng của truyền thông di động ($C_c$) và truyền thông D2D ($C_d$):

$$C_{total} = C_c + C_d \tag{1}$$

Trong đó, $C_c$ và $C_d$ được tính như sau:

$$C_c = \sum_{i=1}^{K} \log_2\left(1+SINR_{ci}\right) + \sum_{j=1}^{M-K} \log_2\left(1+SINR_{cj}\right) \tag{2}$$

$$C_d = K \log_2\left(1+SINR_d\right) \tag{3}$$

Với, $SINR_{ci}$ là SINR của $CUE_i$ chia sẻ tài nguyên với người dùng D2D, $SINR_{cj}$ là SINR của $CUE_j$ không có nhiễu với truyền thông D2D và $SINR_d$ là SINR của